

Approximate Recall Confidence Intervals

WILLIAM WEBBER, The University of Maryland¹

Recall, the proportion of relevant documents retrieved, is an important measure of a effectiveness in information retrieval, particularly in the legal, patent, and medical domains. For large document sets, recall can be estimated by assessing a random sample of documents for relevance; but a measure of the reliability of this estimate is also required. In this article, we examine several methods of calculating confidence intervals on recall estimates. We find that the normal approximation in current use provides poor coverage in many circumstances, even when adjusted to correct its inappropriate symmetry. Analytic and Bayesian methods based on the ratio of binomials are generally more accurate, but perform poorly on small populations. The recommended method derives beta-binomial posteriors on retrieved and unretrieved yield, with fixed hyperparameters, and a Monte Carlo estimate of the posterior distribution of recall. We demonstrate that this method gives mean coverage at or near the nominal level for differing scenarios, while being balanced and stable. We offer advice on sampling design, including the allocation of assessments to the retrieved and unretrieved segments, and compare the proposed beta-binomial with the officially reported normal intervals for recent TREC Legal Track iterations.

Categories and Subject Descriptors: G.3 [Mathematics of Computing]: Probability and Statistics—*experimental design*; G.3 [Mathematics of Computing]: Probability and Statistics—*distribution functions*; H.3.4 [Information Storage and Retrieval]: Systems and Software—*performance evaluation (efficiency and effectiveness)*

General Terms: Measurements, Experimentation, Verification

Additional Key Words and Phrases: Posterior distributions, probabilistic models

ACM Reference Format:

Webber, W., Approximate recall confidence intervals ACM Trans. Inf. Syst. V, N, Article A (January YYYY), 39 pages.

DOI = 10.1145/0000000.0000000 <http://doi.acm.org/10.1145/0000000.0000000>

1. INTRODUCTION

There are many ways to measure the effectiveness of a set-based document retrieval. One of these is recall, the proportion of the relevant documents in the source corpus that are retrieved. Recall is important for domains that aim at comprehensive retrieval, including patent search, medical literature reviews, and document discovery for civil litigation (or *e-discovery*). To measure recall, however, we must know the number of relevant documents, or *yield*, of the corpus. Corpora are generally too large for exhaustive relevance assessment, which would in any case make the retrieval redundant in practical applications. In large retrievals, exhaustive assessment even of the

¹Work performed in part while author was at the University of Melbourne.

This work was supported in part by the Australian Research Council and by NSF award IIS-1065250. Author's address: College of Information Studies, University of Maryland, College Park, MD 20742, The United States of America. Email: wew@umd.edu.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies show this notice on the first page or initial screen of a display along with the full citation. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, to redistribute to lists, or to use any component of this work in other works requires prior specific permission and/or a fee. Permissions may be requested from Publications Dept., ACM, Inc., 2 Penn Plaza, Suite 701, New York, NY 10121-0701 USA, fax +1 (212) 869-0481, or permissions@acm.org.

© YYYY ACM 1046-8188/YYYY/01-ARTA \$10.00

DOI 10.1145/0000000.0000000 <http://doi.acm.org/10.1145/0000000.0000000>

retrieved documents (the retrieved *segment*) may not be feasible. The evaluator must instead turn to estimates.

The yields of the retrieved and unretrieved segments of the corpus can be estimated by drawing a random sample of documents from each segment and assessing them for relevance. The proportion of relevant documents, or *prevalence*, in the sample gives a point estimate of prevalence in the segment, and hence of yield and recall (Section 2.1). Retrieved and unretrieved segments can be sampled at different rates, and each segment may be stratified, to leverage auxiliary evidence about stratum prevalence (Section 2.2).

Besides point estimates of yield and recall, we require a measure of the reliability of these estimates, given sampling error.² One measure of an estimate's reliability is a confidence interval, giving a lower and upper bound which contain the true value with a certain degree of confidence, expressed as a percentage. For instance, one might state that a recall estimate of 0.42 has a 95% confidence interval of 0.31 to 0.55. An *exact* interval of (say) 95% guarantees that for any given true parameter, 95% of re-samples and re-calculations would generate intervals covering that parameter. Exact intervals are generally constructed from the sampling distribution of the statistic of interest. Exact intervals, however, may be unavailable, or too complex to compute or analyze, in which case *approximate* intervals may be used instead [Smithson 2002]. Moreover, for discrete statistics, exact intervals are generally conservative, providing mean coverage above the nominal level (Section 2.5). In contrast, approximate intervals may provide nominal coverage on average, albeit at the expense of under-covering certain parameters, and so may be preferred, even when an exact interval is available (Section 2.6). This article considers approximate intervals.

We examine several methods of calculating confidence intervals on recall in Section 3. One method, widely used to bound the sensitivity of diagnostic tests, treats recall as a simple binomial proportion on relevant documents (Section 3.1) [Simel et al. 1991]. However, it assumes equal sampling over the relevant set, and gives unreliable results when this assumption is violated. A second method, which we term the normal MLE method, uses a normal approximation, a maximum-likelihood estimate (MLE) of variance, and the propagation of error (Section 3.2). This method has been used in e-discovery [Oard et al. 2008]. It frequently under-covers recall (provides coverage less than nominal), sometimes substantially. The remaining methods we consider are new.

Where prevalence is low, the normal MLE method understates uncertainty on yield, and incorrectly assumes a symmetric interval. Adjusting the normal approximation, by adding one or two to the positive and negative sample counts, gives a superior binomial confidence interval [Agresti and Coull 1998; Greenland 2001]. We apply a similar adjustment for the normal approximation of the recall interval (Section 3.3). The adjustment improves coverage in some circumstances, but has mixed overall accuracy.

The normal method assumes that recall is approximately normal in its sampling distribution (the distribution of estimates that occurs on repeated sampling from the one population), which is not in general the case (Section 2.4). A closer approximation is to model recall as (a function of) the ratio between two binomial variables. We adapt an interval on a binomial ratio [Koopman 1984] to bound recall in Section 3.4. The method is generally accurate, but provides no finite population adjustment, and so overstates the interval where sample size is a substantial proportion of a segment.

The previous methods derive intervals from the sampling distribution of the statistic, such as sample yield. An alternative, Bayesian approach is to directly infer a posterior distribution over the population parameter (Section 2.7), and take the interval from the posterior's quantiles. Rather than directly inferring a recall distribution, we

²Assessment error is another, and potentially greater, source of unreliability; see Section 5.1 for a discussion.

infer posterior distributions over the yield of the retrieved and unretrieved segments, and generate from these a Monte Carlo estimate of the distribution of recall, from the percentiles of which a confidence interval is read [Buckland 1984; Chen and Shao 1999].

A beta posterior based on the Jeffreys prior, which sets the Beta distribution hyperparameters α and β to 0.5, provides good coverage for a binomial proportion [Brown et al. 2001] (Section 2.6). Applying beta posteriors to the retrieved and unretrieved segments also gives good general coverage of recall (Section 3.5). However, even with a finite population adjustment, the beta posteriors perform poorly when sample size is a substantial proportion of the population (as is common for the retrieved segment in e-discovery), since they fail to account for the dependence between what is seen in the sample and what is left in the unsampled population.

Sampling without replacement from a finite binomial population is most precisely modelled by a hypergeometric distribution. The conjugate distribution (Section 2.7) to the hypergeometric is the beta-binomial. Our final, and recommended, approach derives beta-binomial posteriors on the retrieved and unretrieved segments (Section 3.6). This method gives the best accuracy of all those considered, and is robust to the finite population case. The information-theoretic most conservative priors (Section 2.8) give better balance than the uniform prior; however, a simple prior of $\alpha = \beta = 0.5$, though not theoretically justified here as for the Jeffreys prior to the binomial, gives best overall results.

The proposed methods for calculating recall confidence intervals are evaluated in Section 4. We assess the performance of the intervals against three criteria. First, the mean coverage of an interval should be close to the nominal; second, the standard error of the coverage should be low; and third, an uncovered true parameter value should be as likely to fall below as above the interval. We also check interval width, to guard against certain degenerate cases that meet the above criteria without providing informative intervals. These criteria must be evaluated over true parameter distributions, and we define three such distributions or scenarios: a broad neutral one; an emulation of an e-discovery environment; and one that explores the finite population case of a small population and large sample (Section 4.1).

Having established the beta-binomial with $\alpha = \beta = 0.5$ as the most reliable method, we examine questions of sample size in Section 4.3, such as the interval-minimizing allocation of assessments to retrieved and unretrieved segments, and the effect of increased sample size on interval width. Finally, we calculate recall intervals on the participants in the Interactive Task of the TREC Legal Track, and compare these intervals with those officially reported for the track, which were generated using the normal-approximation method (Section 4.4).

2. PRELIMINARIES

In this section, we set out preliminary materials, used in Section 3 to design recall confidence intervals. We begin by deriving a point estimate of recall (Section 2.1), and extending it to stratified sampling (Section 2.2). The distributions used to construct recall confidence intervals are the hypergeometric, binomial, and normal; these are introduced in Section 2.3. The sampling distribution of recall itself is considered in Section 2.4, and found to be highly non-normal for a representative sampling scenario. Moreover, the estimator described in Section 2.1 is shown to be biased, and severely so for low prevalence and small sample size in the unretrieved segment. Section 2.5 defines exact confidence intervals, while Section 2.6 describes approximate confidence intervals, and why they are often preferable to their exact counterparts. Bayesian prior and posterior distributions are introduced in Section 2.7. In Section 2.8, we examine the beta-binomial as conjugate prior to the hypergeometric distribution; this forms

Table I. Notation.

Symbol	Meaning	Notes
N_*	Size of corpus	
N_1	Size of retrieved segment	
N_0	Size of unretrieved segment	$N_* = N_1 + N_0$
R_*, R_1, R_0	Number of relevant documents (yield) in above populations	$R_* = R_1 + R_0$
π_*, π_1, π_2	Prevalence of relevant documents in population	$\pi = R_*/N_*$
n_*, n_1, n_0	Size of samples	$n_* = n_0 + n_1$
r_*, r_1, r_0	Number of relevant documents (yield) in sample	$r_* = r_1 + r_0$
p_*, p_1, p_2	Prevalence of relevant documents in sample	$p_* = r_*/n_*$
N, n, R, \dots	Population size, sample size, population yield etc. for variable or unspecified segments	
N_s, n_s, R_s, \dots	Population size, sample size etc. for stratum s (Section 2.2)	

Table II. Document counts in a retrieval.

	Relevant	Not relevant	
Retrieved	R_1	$N_1 - R_1$	N_1
Not retrieved	R_0	$N_0 - R_0$	N_0
	R_*	$N_* - R_*$	N_*

the basis of our most accurate interval estimator. The estimation of intervals using Monte Carlo simulations is described in Section 2.9. Finally, Section 2.10 introduces the propagation of error, used for normal intervals, and derives a propagation of error expression for recall.

2.1. Estimating recall

A set-based retrieval process returns from a corpus those documents which the process estimates to be relevant to a topic. Such a retrieval can also be regarded as a binary classification of corpus documents into relevant and irrelevant classes. Let the number of documents in the corpus be N_* , the number in the retrieved segment be N_1 , and the number in the unretrieved segment be $N_0 = N_* - N_1$. Some R_* of documents in the corpus are actually relevant to the topic; R_1 of these fall in the retrieved segment, and R_0 in the unretrieved segment. Our notation is laid in Table I, and the document counts are summarized in Table II. We refer to the number of relevant documents in a set as that set's *yield*. The recall of a retrieval is the proportion of relevant documents retrieved:

$$\text{Rec} = \frac{R_1}{R_*} = \frac{R_1}{R_1 + R_0} \quad (1)$$

Generally, corpus yield is not known in advance, and the corpus is too large to determine it by exhaustive assessment (which would, in any case, make the retrieval redundant). Estimation based on random sampling is a solution. A size- n simple random sample from a population of N elements is one in which each subset of n items has the same probability of $1/\binom{N}{n}$ of being sampled, and therefore each element has n/N probability of inclusion in the sample [Särndal et al. 1992]. A common mental model of such a sample is drawing n items randomly in sequence from the population without replacement. To estimate recall, we draw a simple random sample of documents from the retrieved and unretrieved segments; assess these documents for relevance; and use

the assessed sample to estimate yield in each segment. Let n_1 and n_0 be the number sampled from the retrieved and from the unretrieved segment, and r_1 and r_0 be the number relevant in the respective samples. The prevalence of relevant documents in the retrieved sample, p_1 , is an estimator of prevalence in the retrieved population:

$$p_1 = \hat{\pi}_1 = r_1/n_1 ,$$

and likewise for the unretrieved sample prevalence, p_0 , on the unretrieved segment. Then, estimators for the yields of the each segment are:

$$\hat{R}_1 = N_1 \hat{\pi}_1 \quad ; \quad \hat{R}_0 = N_0 \hat{\pi}_0 ;$$

and an estimator for recall is:

$$\widehat{\text{Rec}} = \frac{\hat{R}_1}{\hat{R}_1 + \hat{R}_0} = \frac{\hat{R}_1}{\hat{R}_*} . \quad (2)$$

There is an obvious correlation between the numerating sample variable \hat{R}_1 and the denominating sample variable $\hat{R}_* = \hat{R}_1 + \hat{R}_0$, since the latter includes the former; such correlation makes estimation of variance more complicated (Section 2.10). The correlation can be avoided by rewriting the recall estimator as:

$$\widehat{\text{Rec}} = \frac{1}{1 + \hat{R}_0/\hat{R}_1} . \quad (3)$$

As an estimation of a function of a ratio, the recall estimator is biased; that is, $\mathbb{E}[\widehat{\text{Rec}}] \neq \text{Rec}$, the expectation of the estimator does not equal true recall [Hartley and Ross 1954; Cochran 1977]. For mediate prevalences and large samples, the bias is negligible, but for extreme prevalences and small sample sizes, it can be severe; we examine this issue in Section 2.4.

2.2. Stratified sampling

The retrieved and the unretrieved segment can be subdivided into disjoint strata, from each of which a simple random sample is separately drawn. (We reserve the term “segment” to denote the set of documents retrieved or unretrieved by a system; thus, when estimating the recall of a single retrieval, there are two and only two segments.) If prevalence differs systematically between strata, then stratification improves estimation accuracy; and accuracy may be further improved by allocating more samples to strata with high expected variance in their yield estimate [Thompson 2002]. Stratification may be performed for the sake of this improved accuracy, or else it may be a by-product of the retrieval task; for instance, if several (set-based) retrievals are being assessed over the one corpus and task, then the intersections between the retrievals form natural strata [Tomlinson et al. 2007].

As with the retrieved or unretrieved segment, the estimated yield R_s of stratum s is derived from the size of the stratum N_s and the proportion of the stratum sample r_s/n_s that is relevant:

$$\hat{R}_s = N_s r_s / n_s . \quad (4)$$

The estimated yield of the segment T is the sum of the estimated yields of the strata in the segment:

$$\hat{R}_T = \sum_{s \in T} \hat{R}_s , \quad (5)$$

and recall is estimated using these segment estimates in Equation 1. This article focuses upon recall interval estimation under segment simple random sampling case;

extensions to stratified sampling are noted where applicable, and the general case of unequal sampling is discussed in Section 5.1.

2.3. Hypergeometric, binomial, and normal distributions

A corpus can be viewed as a binomial population of relevant and irrelevant documents. If we sample n elements without replacement from a population of size N , in which R elements are positive, then the probability that the number r of positive elements in the sample takes on a particular value k is given by the hypergeometric distribution:

$$\Pr(r = k|N, R, n) = \frac{\binom{R}{k} \binom{N-R}{n-k}}{\binom{N}{n}} \quad (6)$$

where $\binom{a}{b} = a!/(b!(a-b)!)$. The hypergeometric distribution is so named because its successive terms in k follow a hypergeometric series; that is, a series in which $f(k+1)/f(k) = g(k)$, where $g(\cdot)$ is some “simple” function of k . The prefix “hyper” comes from viewing this as an extension of the geometric series, where $g(\cdot)$ is a constant. Setting $g(k) = k$ and $f(0) = 1$ gives the exponential series. (See Dutka [1984] for an historical perspective.) For the hypergeometric distribution, we have:

$$g(k) = \frac{(R-k)(n-k)}{(k+1)(N-R-n+k+1)} \quad (7)$$

As the size of the population N increases, the change that each without-replacement draw makes upon the original population prevalence $\pi = R/N$ diminishes, and the probability of sampling k positive elements is approximated with increasing closeness by the binomial distribution:

$$\Pr(r = k|\pi, n) \approx \binom{n}{k} \pi^k (1-\pi)^{n-k} . \quad (8)$$

The closeness of the approximation depends on the population size N , the sample size n , and the prevalence π . The approximation is weaker for smaller populations, larger samples, and prevalences more divergent from 0.5.

For large n and π not close to 0 or 1, the binomial distribution is in turn approximated by (a discretized version of) the normal distribution:

$$\Pr(r = k|\pi, n) \approx \mathcal{N}(n\pi, \sqrt{n\pi(1-\pi)}) , \quad (9)$$

where \mathcal{N} is the normal distribution function.

Figure 1 compares the hypergeometric, binomial, and discretized normal distributions, for the same population prevalence $R/N = \pi = 0.3$, but different population and sample sizes, N and n . The hypergeometric gives the precise without-replacement sampling distribution. Where n approaches N (left), the binomial and normal approximations overstate the probability of sample prevalence p diverging from population prevalence π . For small n and $N \gg n$ (middle), the binomial approximation is close to the hypergeometric; the symmetry of the normal approximation, however, poorly fits the asymmetry of the hypergeometric, thus overstating the probability of a sample prevalence being towards 0.5. Increasing n , while holding n/N fixed (right), brings the normal approximation closer to the hypergeometric [Feller 1945; Nicholson 1956].

2.4. The sampling distribution and bias of the recall estimator

The hypergeometric, binomial, and normal distributions can be used to model sampling from the retrieved and unretrieved segments of the corpus. Combining them analytically to determine a closed-form expression for the sampling distribution of recall

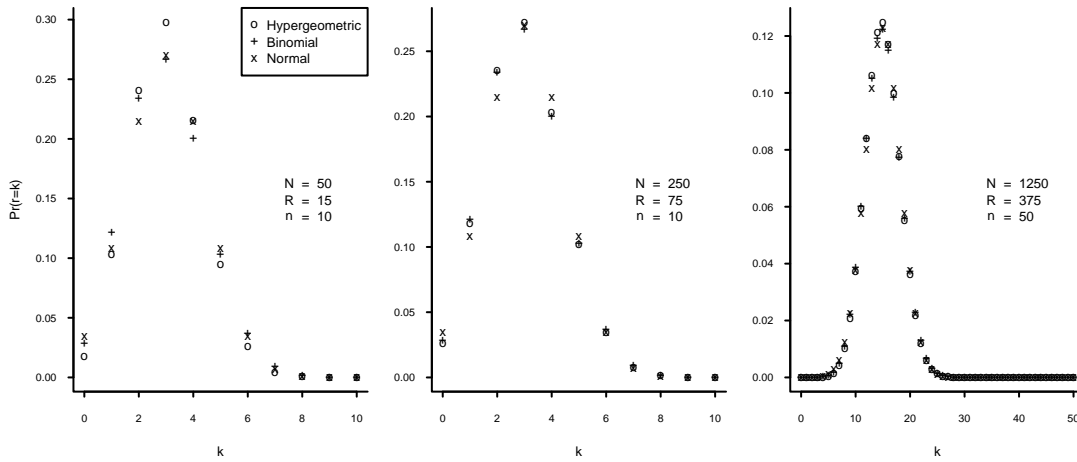


Fig. 1. Probability of drawing k positive instances in n without-replacement draws from a population with N elements, R of them positive, as calculated by the hypergeometric distribution and by the binomial and normal approximations, for different n , N , and R . The normal approximation uses continuity correction; that is, $\Pr(r = k) = \Pr(r < k + 0.5) - \Pr(r < k - 0.5)$.

Table III. Example scenario. The retrieval has 0.5 precision and 0.25 recall. The sampling distribution for this scenario is shown in Figure 2.

Segment	Yield	Population size	Sample size
Retrieved	1,000	2,000	100
Unretrieved	3,000	100,000	100
Total	4,000	102,000	200

is not straightforward; but the distribution for any particular population and sample size can be calculated by brute force means. Figure 2 shows such a distribution for a representative scenario, detailed in Table III. The small number of discrete values that r_0 can take on, combined with the skew of recall's estimator (Equation 2), produces a complex, skewed, multi-modal distribution, having wide gaps around high recall estimates. The normal distribution, also shown in Figure 2, is a poor approximation. These characteristics alert us that recall inferences based upon normal or other approximations may be inaccurate.

The estimator whose sampling distribution is shown in Figure 2 also has a strong positive bias. True recall for this scenario is 0.25, but the mean of the estimator is 0.31. The bias is the result of the skew in the estimator, clearly visible in Figure 2; this is one occasion in which the bias in the estimation of a ratio is non-negligible. This bias is likely to affect the balance and coverage of interval estimators that derive a symmetric bound around the point estimate. Figure 3 shows what happens to estimator bias as prevalence in the unretrieved segment and the size of the unretrieved sample vary. The lower prevalence, and the smaller the sample size, the greater is the bias of the estimator; conversely, higher recall values result in less bias. The bias of the point estimator underlines the importance of employing a reliable confidence interval when describing evaluation results.

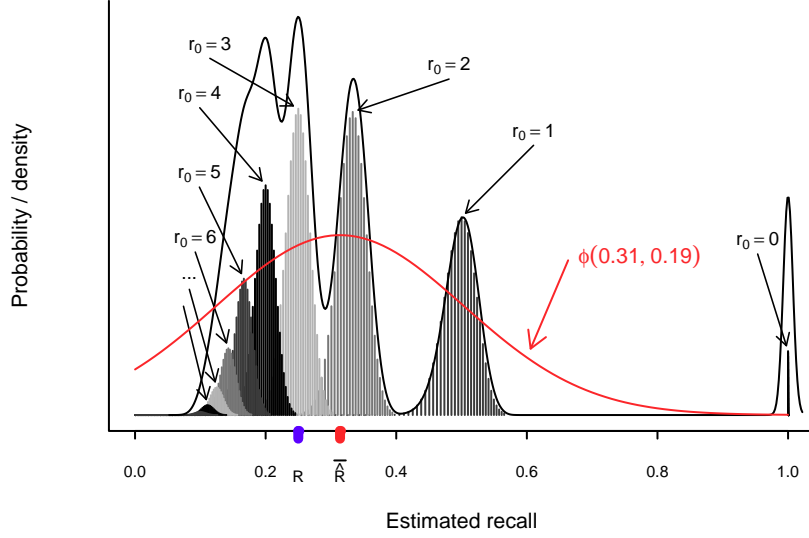


Fig. 2. Sampling distribution of recall for the example scenario shown in Table III. The bars show the sampling distributions induced by r_1 , for different sample values of r_0 . The aggregated probabilities are approximated by a Gaussian kernel density estimate. The density has been scaled visually to fit the probability bars; the height of the density is determined by both the height and the proximity of the bars. Note that r_0 produces a point density at $\text{Rec} = 1$; the density here is for comparison of area only. The true population recall, R , and the mean of the recall estimator, \hat{R} , are marked; the estimator has a strong positive bias. A normal curve with the mean and standard deviation of the sampling distribution is also shown.

2.5. Exact confidence intervals

A common expression of our degree of certainty about the true value of a population parameter θ is a confidence interval. Let $C(X)$ be a function that takes a random sample X and produces an interval $[\underline{\theta}, \bar{\theta}]$, and let P_θ be the probability distribution over samples X given a true parameter value θ . Then $C(x)$ (where x is a realization of X) is a $100 * (1 - \alpha)$ confidence interval if:

$$P_\theta\{\theta \in C(X)\} \geq 1 - \alpha, \quad (10)$$

for all possible values of θ [Lehmann and Romano 2005]. We refer to $P_\theta\{\theta \in C(X)\}$ as the *coverage* of θ by the confidence interval.

The definition of a confidence interval in Equation 10 requires some unpacking to understand. The common interpretation—that a 95% confidence interval on θ means there is a 95% probability that θ lies within the stated interval—is inexact. In the frequentist understanding, θ is fixed, not random, and has no probabilities directly associated with it; it either lies in the interval, or it does not. In Equation 10, what is random is the sample, X ; and the equation can be understood as saying that, whatever the actual value of θ , if we repeatedly drew random samples and calculated the confidence interval $C(\cdot)$ each time, then (in the limit) at least 95% of these confidence intervals would cover θ .

Confidence intervals are often built from a pair of inverted, one-tailed hypothesis tests. Let $S_l(x; \alpha/2)$ be the interval $[\theta_l, \infty]$ of null hypotheses θ_0 that we accept at significance level $\alpha/2$, for the observed sample statistic x , in a one-tailed hypothesis with a lower-tailed alternative; and let $S_u(x; \alpha/2)$ be the corresponding upper-tailed interval $[\infty, \theta_u]$. Then a $1 - \alpha$ confidence interval is defined as $S_l(x; \alpha/2) \cap S_u(x; \alpha/2) = [\theta_l, \theta_u]$.

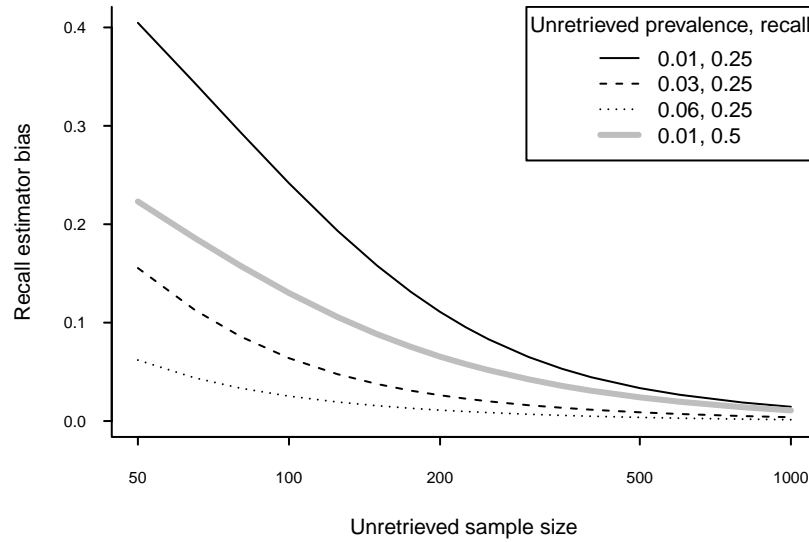


Fig. 3. Bias of recall estimator for scenario described in Table III but with varying unretrieved sample sizes (x axis) and unretrieved yield (and therefore prevalence). Recall and precision are held fixed in the first three cases by scaling retrieved population size and yield by the same amount as unretrieved yield; in the fourth case, retrieved yield and population size is held fixed, and recall increases with the decrease in unretrieved yield. Note the log scale on the x axis.

Informally, the lower end of the confidence interval is the smallest value of θ that a lower-tailed hypothesis test fails to reject, given the observed statistic x , at level $\alpha/2$, and conversely for the upper end. We refer to an interval constructed by inverted hypothesis tests, and which therefore guarantees the coverage expressed in Equation 10, as an *exact* confidence interval.

We say that a confidence interval is a tight one if $P_\theta\{\theta \in C(X)\} = 1 - \alpha$ for all θ . Exact intervals are not always tight ones, and exact intervals on discrete sample distributions such as the binomial generally are not [Neyman 1935]. Therefore, in order to guarantee that an interval derived from a discrete sample variable provides coverage of at least $1 - \alpha$ for each θ , mean coverage will generally be greater than $1 - \alpha$. In other words, for every coverage to be exact, mean coverage must be conservative.

An example of an exact interval on a discrete sample variable is the Clopper-Pearson exact binomial confidence interval on a binomial population, formed from a pair of inverted, one-tailed binomial significance tests [Clopper and Pearson 1934]. Figure 4 shows the coverage of the Clopper-Pearson interval, for a confidence level of 95%, across different population prevalences, for a sample size of 20. The interval guarantees coverage of at least $1 - \alpha$ for all θ , but does not give tight coverage. In fact, the average coverage is 97.7%, and higher still for unbalanced proportions.

2.6. Approximate confidence intervals

The conservatism of exact confidence intervals for discrete distributions has led to interest in approximate intervals on such distributions [Agresti and Coull 1998]. Whereas exact intervals guarantee minimum nominal coverage, approximate intervals seek average coverage at the nominal level (as well as being, in some cases, easier to calculate or analyze). Liu and Kott [2009] use the term “coverage intervals” to de-

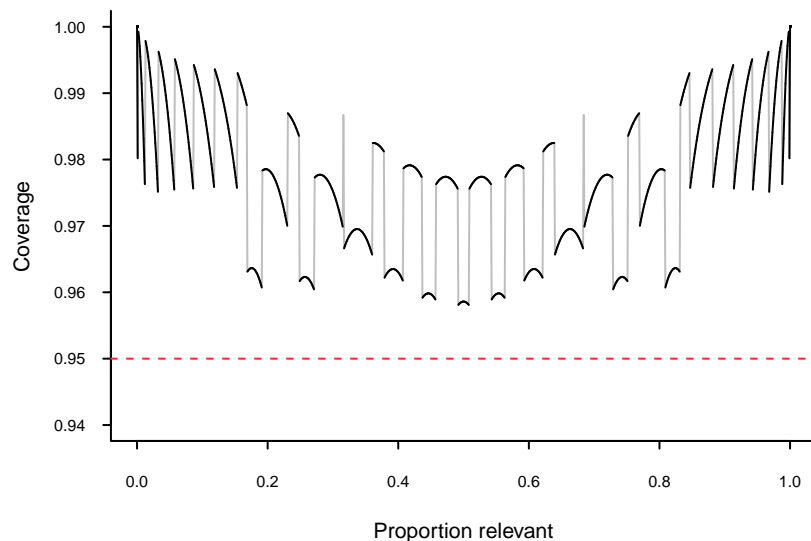


Fig. 4. Coverage of the exact binomial confidence interval, with a nominal confidence level of 95%, for a sample size of 20, across different population prevalences. Discontinuities are shown in gray.

scribe intervals aiming at mean, rather than minimal, nominal coverage; however, we will continue to refer to them as (approximate) confidence intervals.

A common approximate interval for the binomial, the Wald interval, calculates the upper and lower bounds as quantiles of a normal distribution centered upon the sample proportion, which is equivalent to inverting two one-sided normal tests of sample standard error [Agresti and Coull 1998]. If a size n sample produces sample prevalence $p = r/n$, then an estimate of sample variance is:

$$\widehat{\text{Var}}_p = \frac{p(1-p)}{n}. \quad (11)$$

Note that this is actually a biased estimator; the unbiased estimate has $(n-1)$ as its divisor [Cochran 1977, Chapter 3]; however, the variant given in Equation 11 is in almost universal use, and so is the one employed throughout this article. The standard error s is the square root of this variance, and the confidence interval is $p \pm s \cdot z_{\alpha/2}$, where $z_{\alpha/2}$ is the $\alpha/2$ quantile of the standard normal cumulative distribution function (for instance, $z_{0.05/2} = 1.96$).

The Wald interval provides unsatisfactory coverage, particularly for unbalanced population proportions, as illustrated in Figure 5. Mean coverage for this example is 85.1%, and coverage drops close to 0 for edge cases. The correct interval is asymmetric when $p \neq 0.5$, since a hypothesized lower bound of (say) 0.07 has a lower standard error than a hypothesized upper bound of (say) 0.15. The Wald interval, however, is always, incorrectly, symmetric. As an extreme case, consider a sample proportion $p = 0$. The Wald method produces a zero-width, zero-centered interval, even though a population with a non-zero prevalence can still produce a zero prevalence sample.

An alternative approximate binomial interval is the Wilson (or score) interval. This interval also inverts normal hypothesis tests, but uses the prevalences at the candidate lower and upper bounds to calculate test standard errors. Thus, the lower half of the interval is shorter than the upper for $p < 0.5$, and vice versa for $p > 0.5$. The precise

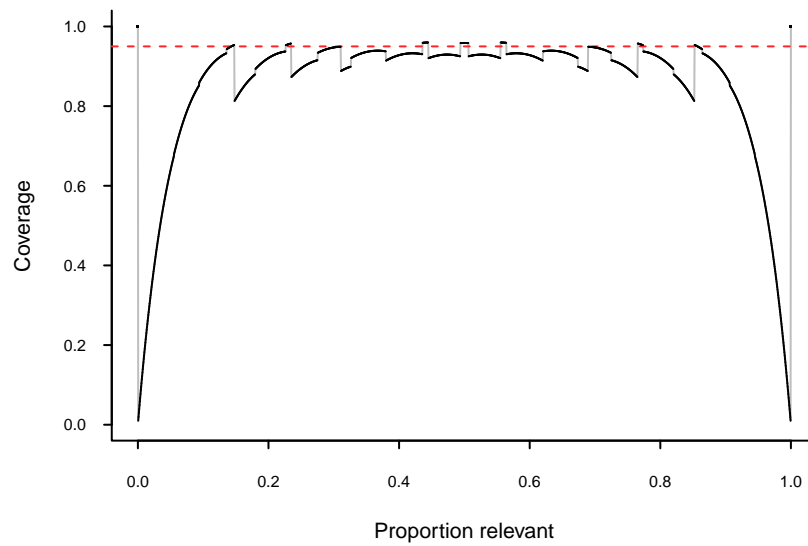


Fig. 5. Coverage of the Wald binomial confidence interval, with a nominal confidence level of 95%, for a sample size of 20, across different population prevalences. Discontinuities are shown in gray.

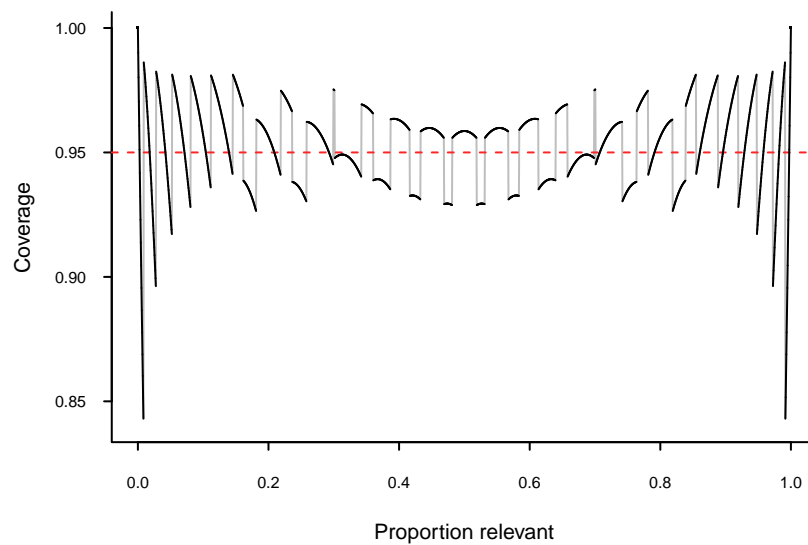


Fig. 6. Coverage of the Wilson binomial confidence interval, with a nominal confidence level of 95%, for a sample size of 20, across different population prevalences. Discontinuities are shown in gray.

interval is:

$$\frac{p + (z_{\alpha/2}^2/2n) \pm z_{\alpha/2} \sqrt{[p(1-p) + z_{\alpha/2}^2/4n]/n}}{1 + z_{\alpha/2}^2/n} \quad (12)$$

[Agresti and Coull 1998]. The coverage of the Wilson interval is shown in Figure 6. Mean coverage here is 95.3%.

Agresti and Coull [1998] show that a 95% Wilson confidence interval is closely approximated by adding two to the count of positive and of negative items observed in the sample to form an adjusted sample size \tilde{n} and sample prevalence \tilde{p} , and using the adjusted \tilde{p} and \tilde{n} in the Wald interval, to produce an interval of:

$$\tilde{p} \pm z_{0.025} \sqrt{\tilde{p}(1-\tilde{p})/\tilde{n}}. \quad (13)$$

More generally, we add $z_{\alpha/2}^2/2$ positive and $z_{\alpha/2}^2/2$ negative observations, where $z_{\alpha/2}$ is the appropriate quantile of the standard cumulative normal distribution (for instance 1.96 for a 95% interval). We revisit adding dummy observations for recall confidence intervals in Sections 3.5 and 3.6.

The above discussion has considered two-sided confidence intervals for a binomial proportion. Cai [2005] observes that methods for interval estimation have different characteristics for one-side confidence intervals (which set only a lower or only an upper bound), since errors at one end of the interval cannot be compensated for by errors at the other. The Wilson interval, while giving good coverage for two-sided intervals, is systematically biased in the true proportion p for one-sided intervals. Cai [2005] instead corrects the Wald interval with higher terms from the Edgeworth expansion (extending the method of Hall [1982]). Liu and Kott [2009] survey one-sided intervals for binomial proportions, including under more complex (for example, stratified) sampling schemes. In this article, we consider only two-sided confidence intervals on recall.

2.7. Prior and posterior distributions

Section 2.5 set out the frequentist understanding of the confidence interval. The true parameter (for us, recall) is fixed; what is variable is the sample. A single sample has been observed; we reason from this sample to the notional result of like samples and the intervals derived from them which, as they approach infinity in number, will cover the true parameter with the frequency specified by the interval's confidence level. The reasoning itself takes the form of (for a two-tailed interval) a pair of hypotheses about what the true parameter might be, and how likely the observed sample would be under each hypothesis; these hypotheses bound the interval.

An alternative approach is offered by the Bayesian viewpoint. Under this understanding, the true parameter is treated as a random variable, the randomness coming from our subjective state of incomplete knowledge about its value. Our statistical reasoning takes the form of inferring a probability distribution for this parameter. Estimates and decisions are derived from the inferred distribution over the true parameter [Gelman et al. 2004]. In the case of confidence intervals, a $1 - \alpha$ interval is derived simply by reading off the $\alpha/2$ and $1 - (\alpha/2)$ quantiles of the distribution's cumulative distribution function. (Bayesian statistics describes an alternative formalism for a confidence interval called a credible interval, which supports the intuitive understanding of the true parameter falling within the stated interval with a certain probability; there is little difference in application, particularly for approximate intervals evaluated in terms of their coverage [Bolstad 2007].)

As attractive as the Bayesian approach is, it relies crucially on the positing of a prior distribution over the parameter; that is, a distribution which lays out our subjective

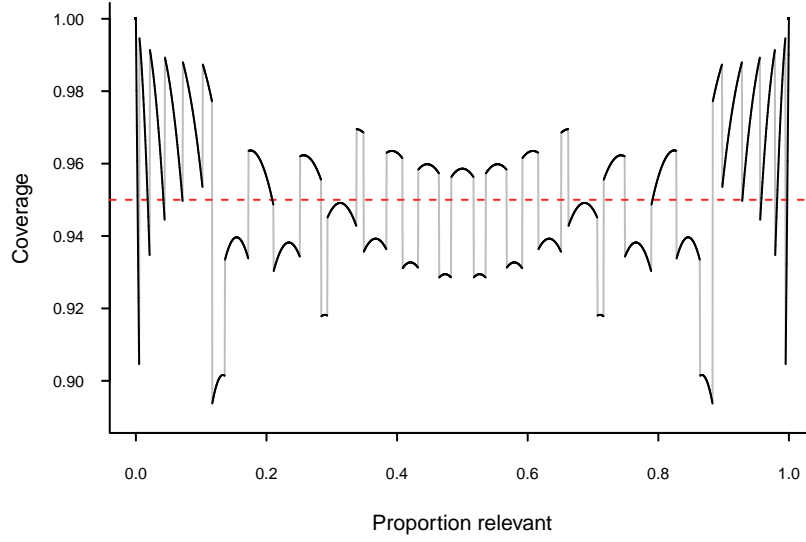


Fig. 7. Coverage of the binomial confidence interval based upon a Jeffreys prior, with a nominal confidence level of 95%, for a sample size of 20, across different population prevalences. Discontinuities are shown in gray.

belief about the parameter's likely values, prior to observing the evidence of the sample. This prior distribution is then updated by the observed evidence to form a posterior distribution; and it is from this posterior distribution that estimates and decisions are derived.

Let θ be the parameter of interest, and x the evidence we have observed (for instance, from a random sample). We wish to infer $p(\theta|x)$, a probability distribution for θ , given the observed evidence, x . Using Bayes' rule, the expression is re-written as:

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} = \frac{p(x|\theta)p(\theta)}{\int p(x|\theta)p(\theta) d\theta} . \quad (14)$$

That is, the posterior probability of θ given x is the prior probability of θ , times the likelihood of x given θ , divided by the marginal probability of x , derived by integrating over our prior distribution on θ [Gelman et al. 2004].

For computational and analytical convenience, it is common to choose the prior distribution $p(\theta)$ from a family of distributions such that the posterior, $p(\theta|x)$, is from the same family, given the distribution of the likelihood, $p(x|\theta)$. We say in this case that the distribution of the prior is conjugate to that of the likelihood. For instance, the beta distribution, $Beta(\alpha, \beta)$, with probability distribution function:

$$f(q; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} q^{\alpha-1} (1 - q)^{\beta-1} \quad (15)$$

(where $\Gamma(\cdot)$ is the Gamma function; $\Gamma(n) = (n - 1)!$ for positive integer n) is conjugate prior to the Binomial. We use $f(q; \alpha, \beta)$ to express the prior probability $\Pr(\pi = q)$ that the true prevalence π is q . If the sample x , of size n , contains k positive instances, and $n - k$ negative instances, then the prior distribution over π of $Beta(\alpha, \beta)$ is updated to the posterior distribution of $Beta(\alpha + k, \beta + n - k)$.

A particular prior distribution is instantiated from a distribution family by the choice of hyperparameters, such as α and β for the beta distribution. In the absence of

previous information about the parameter of interest, a non-informative prior is generally chosen, one which is non-committal about the value of the parameter. A simple approach for a single population parameter θ of finite range is to regard all values of θ as equally likely, by choosing hyperparameters that give a uniform prior distribution over θ . For the beta prior to the binomial parameter π , this means setting $\alpha = \beta = 1$. The uniform prior, however, may be too strong a prior; for small samples of unbalanced binomial populations, it may pull inferences on π too far towards $\pi = 0.5$.

Various more formal methods have been proposed for deriving non-informative, sufficiently weak priors. One such method, known as the Jeffreys prior, selects a prior distribution proportional to (for a single parameter) the square root of the parameter's expected Fisher information. This formulation is chosen because it is invariant to transformations of model parameters (for instance, if we convert a proportion to a logit scale before inferring probabilities) [Jeffreys 1946].³ (Section 2.8 describes an alternative information-theoretic basis for selecting a non-informative prior.) The Jeffreys prior for the binomial is $Beta(0.5, 0.5)$; after observing k positive and $n - k$ negatives, the resulting posterior is $Beta(0.5 + k, 0.5 + n - k)$. The $\alpha/2$ and $1 - \alpha/2$ quantiles of the posterior provide an approximate $1 - \alpha$ interval on π .⁴ As the Beta is a continuous distribution in the range $[0, 1]$, the lower quantile is always greater than 0, and the upper less than 1, so that these values are never included in the confidence interval, and so have 0 coverage. The confidence interval can be simply adjusted by setting its lower end to 0 if $k = 0$, and the upper end to 1 if $k = n$. Figure 7 shows the coverage of a 95% Jeffreys interval for a sample size of 20; mean coverage here is 95.1%. The Jeffreys and Wilson intervals have been recommended as giving the best two-sided coverage for the binomial [Brown et al. 2001; Agresti and Coull 1998]. The one-sided Wilson interval is systematically biased in p (lower-bound interval under-covers for low p , over-covers for high p); the Jeffreys interval avoid this systematic bias [Liu and Kott 2009].

2.8. Beta-binomial as prior to hypergeometric

Sampling without replacement from a finite (as opposed to infinite) binomial population is most precisely modelled by the hypergeometric distribution. The conjugate prior to the hypergeometric is the beta-binomial. We express the prior probability, $\Pr(R = s)$, that the yield R of a population of size N is s , as $BetaBin(\alpha, \beta; N)$, having the probability distribution function:

$$g_1(s; N, \alpha, \beta) = \binom{N}{s} \frac{B(s + \alpha, N - s + \beta)}{B(\alpha, \beta)}, \quad (16)$$

where $B(x, y)$ is the beta function, $\Gamma(x)\Gamma(y)/\Gamma(x + y)$. Having sampled n elements and observed k positives, the posterior distribution on R is:

$$g_p(s|k; N, n, \alpha, \beta) = \binom{N - n}{s - k} \frac{B(s + \alpha, N - s + \beta)}{B(\alpha + k, \beta + n - k)}, s = k, k + 1, \dots, N - n + k. \quad (17)$$

Note that $g_p(s|k; N, n, \alpha, \beta) = g_1(s - k; N - n, \alpha + k, \beta + n - k)$, establishing conjugacy.

As with the beta prior to the binomial, setting $\alpha = \beta = 1$ gives a uniform prior for R over $[0 \dots N]$; but again, this may be insufficiently weak for small samples from un-

³Fisher information measures the information that observed data gives about the population parameter estimated, in the form of the rate of change in the likelihood function as the population parameter changes; the higher the Fisher information, the more information that an observation has about the parameter, and the lower the mean squared error of the estimator [Lehmann and Casella 1998].

⁴The overloading of α as both the complement of the confidence level, and as a hyper-parameter to the beta (and later beta-binomial) prior, is regrettable, but both usages are too firmly established to be easily neglected. Which usage is meant should be clear by context.

balanced populations. We could weaken the prior by taking the same hyper-parameter settings of $\alpha = \beta = 0.5$ as for the Jeffreys prior to the binomial. The theoretical justification, however, is lacking for the beta-binomial: since the discrete distribution of the yield parameter s is not differentiable, the Fisher information of the prior cannot be calculated [Berger et al. 2008].

An alternative, information-theoretic approach to deriving a prior is described by Dyer and Chiou [1984], in which the *most conservative prior* is the one that maximizes the expected information gain from the observed data, as measured by the Kullback-Leibler divergence between prior and posterior distributions. Another way of putting this is that we wish the likelihood function (Section 2.7) to dominate the prior, by drawing the posterior distribution as close as possible to it in expectation. Let $\underline{\omega}$ be the vector of hyperparameters (α and β for the beta-binomial), and let $g(\theta)$ be the prior and $g(\theta|\cdot)$ the posterior distributions. Then the information gain from observing data x is:

$$I[g(\theta|x;\underline{\omega}), g(\theta;\underline{\omega})] = \int g(\theta|x;\underline{\omega}) \ln \left(\frac{g(\theta|x;\underline{\omega})}{g(\theta;\underline{\omega})} \right) d\theta, \quad (18)$$

in which integration is over the domain of θ . The expected information gain $\bar{D}(\underline{\omega})$ for a vector of hyperparameters ω is the expectation of Equation 18 with respect to the evidence variable X :

$$\bar{D}(\underline{\omega}) = \int m(x;\underline{\omega}) I[g(\theta|x;\underline{\omega}), g(\theta;\underline{\omega})] dx; \quad (19)$$

where $m(x;\underline{\omega})$ is the marginal distribution of X , given the hyperparameters. The integrands in Equation 18 and Equation 19 are replaced by sums for discrete parameters and variables. The most conservative hyperparameters $\underline{\omega}$ are those that maximize $\bar{D}(\underline{\omega})$ in Equation 19.

The beta-binomial prior $BetaBin(\alpha, \beta)$ to the hypergeometric distribution has α and β as its hyperparameters, and is conditional on population size N and sample size n . Dyer and Pierce [1993] show that the average information gain under Equation 19 for this prior is:⁵

$$\begin{aligned} \bar{D}_{BB}(\alpha, \beta; N, n) &= \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\alpha + \beta + N)} \\ &\times \sum_{x=0}^n \sum_{k=x}^{N-n+x} \binom{n}{x} \binom{N-n}{k-x} \Gamma(\alpha + k) \Gamma(\beta + N - k) \\ &\times \ln \left[\frac{\binom{N-n}{k-x} \Gamma(\alpha) \Gamma(\beta) \Gamma(\alpha + \beta + n)}{\binom{N}{k} \Gamma(\alpha + x) \Gamma(\beta + n - x) \Gamma(\alpha + \beta)} \right]. \end{aligned} \quad (20)$$

Dyer and Pierce [1993] also demonstrate that Equation 20 is concave and symmetric in α and β , so the maximum occurs where $\alpha = \beta$. Finding the value $\alpha = \beta$ which maximizes Equation 20 (which we do as an optimization problem) provides the most conservative prior [Dyer and Pierce 1993].

The most conservative prior in Equation 20 depends on both sample size n and population size N . Figure 8 shows most conservative priors for various N and n/N . We conjecture that there is an inflection point around $n/N = 0.8$ and slightly be-

⁵Correcting Equation 3.1.3 of Dyer and Pierce [1993], which has the error $\binom{n}{x}$ instead of $\binom{N}{k}$ in the final line. If referring to said article, note also that the middle equation on Page 2131 should have $\Gamma(\beta + N - k)$ in the numerator, instead of $\Gamma(\beta + N - x)$.

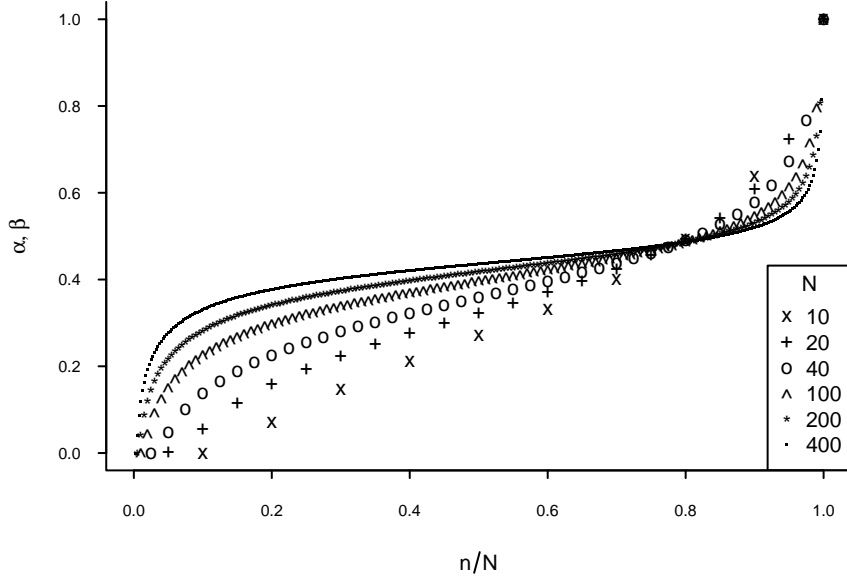


Fig. 8. Most conservative hyperparameters α, β to the beta-binomial as prior to the hypergeometric, for varying N and n .

low $\alpha = \beta = 0.5$, and that for any fixed ratio $C = n/N$, the most conservative values of α and β approach this inflection value as N goes to infinity. Equation 20 is time-consuming to calculate for large N and n , as the double sum gives it complexity $O(Nn)$; our conjecture suggests that approximations are reasonable in such cases. (In the experiments in Section 4, we place an upper bound of 1,000 on N and 800 on n , or $1,000 - (N - n)$ if n is within 200 of N ; that is, in the right margin of Figure 8. With these bounds, the α and β values near the conjectured inflection point are calculated.)

As theoretically attractive as the most conservative prior is, it has anomalous behaviour for extremely small sample sizes. If the sample size $n = 1$, then the most conservative prior collapses to 0, since in this case posterior distribution and standardized likelihood function both have $p(s) = 1$, whether s is 0 or N . In other words, the posterior collapses to a point distribution function, assigning certainty that the number of positives s in the population is either 0 or 1, depending on whether the single sample is negative or positive. Most conservative priors for very small sample sizes likewise tend to be too weak, producing distributions that are too narrow. The potential for this to lead to inaccurate coverage is observed empirically in Section 4.

2.9. Monte Carlo estimation of intervals

If we have a posterior distribution over a population parameter, we can take a $1 - \alpha$ confidence interval from the posterior's $\alpha/2$ and $1 - \alpha/2$ quantiles. These quantiles can be directly calculated from the posterior's inverse cumulative distribution function, if known. If the population parameter of interest is a function of other population parameters, for which we have posteriors, then its distribution follows from the joint distribution of the constituent parameters. For recall, the constituent distributions are the (beta or beta-binomial, say) posteriors over the yield of the retrieved and unretrieved segments. Analytically deriving a closed-form expression for the inverse cumulative density function of this joint distribution, however, is no easy task.

For discrete distributions, it is possible (given time) to calculate the quantiles of the distribution by exhaustive enumeration of parameters and probabilities. For recall, one would calculate the probabilities of all $(N_0 + 1) \cdot (N_1 + 1)$ yield combinations, and the recall for each; sort by recall; and then accumulate probabilities from each end until the desired quantiles are achieved. This naive approach is $O(N^2)$ in both time and memory. An improved algorithm searches the parameter space from each end for successively higher (or lower) recall values, but even this algorithm is time- and memory-intensive for large populations. And such an approach cannot be applied to continuous distributions.

Rather than exhaustively compute the parameter combinations, we can estimate the interval by Monte Carlo samples from the parameter space [Buckland 1984; Chen and Shao 1999]—assuming it is possible efficiently to sample from the distribution of each constituent parameter. For the case of recall, the beta and beta-binomial posteriors on the retrieved and unretrieved segments can be efficiently sampled from [Cheng 1978]. We draw s such samples in pairs, one from the posterior of each segment, and calculate the recall for each sample. These s recall values are sorted, and the $\alpha/2$ and $1 - \alpha/2$ quantiles estimate the like quantiles of the true distribution, and hence estimate the interval. The accuracy of the estimate grows with the number of samples. For bootstrap resampling, a rule-of-thumb of 1,000 replications are suggested as giving sufficient accuracy for confidence interval calculation [Efron and Tibshirani 1993]. For the methods described in Section 3.5 and Section 3.6, we employ 10,000 Monte Carlo draws.

2.10. Propagation of error

Confidence intervals are often approximated using a normal distribution centered around the point estimate, as with the Wald interval. Where the sample statistic is aggregated from component values, the standard error of the normal approximation can also be aggregated from that of the component estimates. If we have a random variable X that is a function of other random variables A and B , written:

$$X = f(A, B) , \quad (21)$$

then, by the theory of *propagation of error*:

$$\text{Var}(X) = \left(\frac{\partial f}{\partial A} \sigma_A \right)^2 + \left(\frac{\partial f}{\partial B} \sigma_B \right)^2 + 2 \left| \frac{\partial f}{\partial A} \frac{\partial f}{\partial B} \right| \text{Cov}_{AB} , \quad (22)$$

where $\sigma_A = \sqrt{\text{Var}(A)}$ is the standard deviation of A , Cov_{AB} is the covariance between A and B , and $|x|$ is the absolute value of x [Taylor 1997].

The estimate of recall in Equation 3 is a function of two variables, \hat{R}_1 and \hat{R}_0 . Since the estimators \hat{R}_0 and \hat{R}_1 are sampled from different document segments, they are independent (unlike \hat{R}_1 and \hat{R}_* in Equation 2), so the covariance term in Equation 22 can be ignored. It can be shown that:

$$\frac{\partial \widehat{\text{Rec}}}{\partial \hat{R}_0} = - \frac{1}{\hat{R}_1 \left(1 + \hat{R}_0 / \hat{R}_1 \right)^2} , \quad (23)$$

and that:

$$\frac{\partial \widehat{\text{Rec}}}{\partial \hat{R}_1} = \frac{\hat{R}_0}{\left(\hat{R}_1 + \hat{R}_0 \right)^2} . \quad (24)$$

Applying Equation 22, we find that:

$$\widehat{\text{Var}}(\widehat{\text{Rec}}) = \frac{\widehat{\text{Var}}(\widehat{R}_1)\widehat{R}_0^2 + \widehat{\text{Var}}(\widehat{R}_0)\widehat{R}_1^2}{(\widehat{R}_1 + \widehat{R}_0)^4}, \quad (25)$$

The values of $\widehat{\text{Var}}(\widehat{R}_1)$ and $\widehat{\text{Var}}(\widehat{R}_0)$ are the usual estimates of binomial sampling variance for their respective segments, given in Equation 11. If stratified sampling is employed, then the yield estimate \widehat{R}_T for a segment T is the sum of the yield estimates \widehat{R}_s for the strata that make up the segment, $s \in T$ (Equation 5). The variance estimate for each segment is likewise summed from the variance estimates for each stratum in that segment [Oard et al. 2008]:

$$\widehat{\text{Var}}(\widehat{R}_T) = \sum_{s \in T} \widehat{\text{Var}}(\widehat{R}_s). \quad (26)$$

A naive approach to calculating the propagation of error views recall as a function of \widehat{R}_1 and \widehat{R}_* (working from Equation 2, instead of Equation 3), and ignores the covariance term in Equation 22. Recall variance is estimated in this way in Oard et al. [2008] (see in particular Equation 21 of that paper). The naive variance estimate essentially includes the variance of \widehat{R}_1 twice, and so overstates the variance of the recall estimate. The degree of overstatement depends upon the size of $\widehat{\text{Var}}(\widehat{R}_1)$ relative to $\widehat{\text{Var}}(\widehat{R}_0)$; roughly speaking, overstatement will be greatest for retrievals that have high recall but low precision, and in which unretrieved yield is low. We revisit this question empirically in Section 4.4.

3. RECALL CONFIDENCE INTERVALS

The previous section established the materials; this section uses them to build recall confidence intervals. We propose nine different intervals from five different families, summarized in Table IV, moving from the most general approaches to the most specific. The first four methods use normal approximations; the fifth, an analytical interval on a ratio between binomials; and the last four, Bayesian segment posteriors (the first the beta prior to the binomial, and the next three beta-binomial priors to the hypergeometric) and Monte Carlo generation. These methods are then evaluated in Section 4, with the last of them found to be the most accurate.

3.1. Sensitivity as a binomial proportion

The sensitivity of a classifier is the proportion of positive instances correctly classified. The formulation is the same as recall, though the connotation is primarily predictive rather than retrospective. In the field of diagnostic medical testing, Simel et al. [1991] treat sensitivity as a binomial proportion on the population of true positives. They then apply an approximate Wald interval to this proportion (Section 2.6). The $1 - \alpha$ confidence interval on the recall estimate is:

$$\widehat{\text{Rec}} \pm z_{\alpha/2} \cdot \sqrt{\widehat{\text{Rec}}(1 - \widehat{\text{Rec}})/n}. \quad (27)$$

where $z_{\alpha/2}$ is the $\alpha/2$ quantile of the normal cumulative density function. We refer to this as the *naive binomial* confidence interval method.

Aside from the poor coverage of the normal approximation to the binomial (Section 2.6), the naive binomial method assumes that all positive instances are equally sampled. This assumption, though valid in the applications considered by Simel et al. [1991], does not generally hold for retrieval recall estimation. In particular, the re-

Table IV. Recall confidence intervals.

Name	Section	Summary	Coverage
Naive binomial	3.1	Treat sensitivity as binomial proportion sampled from true positives.	Very poor
Normal MLE	3.2	Normal approximation with maximum-likelihood variance estimate, aggregated using propagation of error.	Poor
Laplace	3.3	Add one to positive and negative sample counts in each segment, then as per Normal MLE.	Poor
Agresti-Coull	3.3	Add two to sample counts, then as per Normal MLE.	Poor
Koopman	3.4	Monotonic decreasing transform of interval on ratio between binomial variables.	Fair-Good
Beta Jeffreys	3.5	Infer beta posteriors from Jeffreys priors on segments; Monte Carlo generation of recall distribution; take quantiles of generated distribution.	Fair
Beta-binomial Uniform	3.6	Infer beta-binomial posteriors from uniform priors on segments; then as per Beta Jeffreys.	Fair
Beta-binomial MCP	3.6	Infer beta-binomial posteriors from most conservative priors on segments; then as per Beta Jeffreys.	Good
Beta-binomial $\alpha = \beta = 0.5$	3.6	Infer beta-binomial posteriors from priors with hyper-parameters $\alpha = \beta = 0.5$ on segments; then as per Beta Jeffreys.	Best

trieved and untrieved segments are separately sampled, and at different rates. The same divergence in sampling means the method extends poorly to stratified sampling.

3.2. Normal with maximum-likelihood variance

The variance of the recall estimator can be derived from the variances of the segment yield estimators using the theory of propagation of error (Section 2.10, Equation 25). The standard error of the estimator is the square root of the variance:

$$s_R = \sqrt{\widehat{\text{Var}}(\widehat{\text{Rec}})}$$

If we assume that $\widehat{\text{Rec}}$ is normally distributed, and has equal variance along its range, then a $1 - \alpha$ confidence interval is:

$$\widehat{\text{Rec}} \pm z_{\alpha/2} \cdot s_R. \quad (28)$$

Note that this range is centered around the point estimate $\widehat{\text{Rec}}$. Estimates of $\widehat{\text{Var}}(\widehat{R}_1)$ and $\widehat{\text{Var}}(\widehat{R}_0)$ are required for Equation 25. The maximum-likelihood estimate (MLE) is given in Equation 11. We refer to the approximate normal CI with this variance estimate as the Normal-MLE method. The method can be extended to stratified sampling, by summing the variance of each stratum to calculate the variance of the segments containing those strata [Oard et al. 2008].

The Normal-MLE method has two main problems. First, the sampling distribution of recall may diverge from the normal (Section 2.4). At the very least, recall is bounded by the range $[0, 1]$, whereas the normal distribution is unbounded.⁶ And second, the MLE

⁶Bounding the logit of recall using a normal approximation is discussed as future work in Section 5.1.

of $\text{Var}(\widehat{R})$ understates uncertainty for extreme sample prevalences p . The problem is similar to that of the Wald interval (Section 2.6). In particular, if $p_0 = 0$, then $\widehat{\text{Var}}(\widehat{R}_0)$ is 0 and the recall estimate is 1 ± 0 , but perfect recall is certainly not assured.

3.3. Laplace and Agresti-Coull adjustments to the normal

As mentioned in Section 2.6, Agresti and Coull [1998] correct the Wald confidence interval by adding 2 to the count of positive and negative instances in the sample, giving results close to the highly-accurate Wilson interval. A similar idea can be applied to the recall interval, by adding 1 or 2 to the positive and negative sample counts in the retrieved and unretrieved segments. We denote the addition of 2 as the Agresti-Coull adjustment, and of 1 as the Laplace adjustment (from Laplace's prior to the binomial) [Agresti and Caffo 2000; Greenland 2001]. In stratified sampling, the adjustment would be applied to each stratum individually, though the adjustment amount might depend upon the number of strata.

The modified counts give variant sample proportions \tilde{p} , and (through Equations 11 and 25) adjusted estimates of recall standard error, \tilde{s}_R . The modified counts are also used (in Equation 2) to determine the adjusted center-point for the interval range, $\widetilde{\text{Rec}}$ (though the point estimate would remain $\widehat{\text{Rec}}$; the effect of the adjusted center-point is to induce the desired asymmetry of the interval). The adjusted values are used in Equation 28 to calculate the interval on recall. These adjustments make the interval more robust to extreme sample prevalences, but do not account for the non-normality of recall's sampling distribution.

3.4. Ratio between binomial proportions: Koopman

The ratio between two binomial variables is not normally distributed. Koopman [1984] derives an approximate confidence interval on the ratio of two independent binomial proportions, which he refers to as the chi-square method. (The equations are too lengthy to reproduce here; see Section 2.2 of Koopman [1984] for details.) The formula for recall in Equation 3 is a monotonically decreasing function of the ratio between the independent binomial proportions \widehat{R}_0 and \widehat{R}_1 . Therefore, a $1 - \alpha$ confidence interval on the latter ratio will also be a $1 - \alpha$ confidence interval on recall [Smithson 2002]. Interval coverage and balance will be unaffected by the transformation, though width (in parameter space) may have different characteristics (for instance, that the original interval has the minimum width in ratio of proportions does not guarantee that the transformed interval will have minimum width in recall). We refer to this as the Koopman interval. The method does not have a straightforward extension to stratified sampling, since the samples from retrieved and unretrieved segments cease being simple binomial random samples.

3.5. Ratio between beta posteriors with Jeffreys priors

A beta posterior based on a Jeffreys prior provides a binomial confidence interval with good coverage (Section 2.7). An extension of this method can be applied to derive an interval on recall. Beta posteriors are inferred from a Jeffreys prior on the prevalences of the retrieved and unretrieved segments, π_1 and π_0 . A pair of independent observations is repeatedly drawn from these posteriors, for the unsampled part of each segment only (providing a finite-population adjustment), and recall calculated on that pair, to generate a Monte Carlo distribution over the posterior on recall 2.9. A $1 - \alpha$ confidence interval is derived by taking the $\alpha/2$ and $1 - \alpha/2$ quantiles of these recall observations. We refer to this as the Beta Jeffreys recall interval. Stratified sampling is handled by separately deriving posteriors for each stratum, and drawing samples from each such posterior.

As with the naive Jeffreys interval on a binomial proportion (Section 2.7), the quantiles of the Jeffreys (and indeed any beta) posterior on the proportion relevant never extend to 0, even if there are no positive instances in a segment sample, nor 1, even if there are no negative instances. The most salient effect of this for the interval on recall is that the upper limit of the interval will never be strictly 1, rounding effects aside, even if there are no relevant documents in the sample from the unretrieved set. The upper limit will approach 1 for larger sample sizes; but for extremely small samples from the lower stratum, the upper limit may be well short of 1. This situation can be treated as a special case by setting the upper limit to 1 on an all-irrelevant sample from the unretrieved segment; we do not, however, make that adjustment here.

3.6. Ratio between beta-binomial posteriors with uniform or most conservative priors

Both Koopman and Beta Jeffreys intervals are based on the binomial distribution. In practice, though, the document population is finite, and the hypergeometric distribution is the more accurate model, the conjugate prior to which is the beta-binomial (Section 2.8). For our final approximate recall interval, beta-binomial posteriors are inferred for the yields of the retrieved and unretrieved segment, R_1 and R_0 , and a Monte Carlo distribution over recall is generated by drawing paired observations from these posteriors. As with the Beta Jeffreys method, the $\alpha/2$ and $1 - \alpha/2$ quantiles of the generated distribution provide the $1 - \alpha$ confidence interval on recall.

Section 2.8 discussed the selection of the $\alpha = \beta$ hyperparameters to the beta-binomial prior. We consider three choices for these hyperparameters. The first sets $\alpha = \beta = 1$, giving a uniform prior (Beta-binomial Uniform). The second selects the most conservative prior for the population and sample size, as described in Section 2.8 (Beta-binomial MCP). And the third mimics the Jeffreys prior to the binomial, by setting $\alpha = \beta = 0.5$ (Beta-binomial $\alpha = \beta = 0.5$), which is also close to the apparent inflection point for the most conservative prior (Figure 8). The pseudo-Jeffreys hyperparameters do not have the same theoretical foundation in Fisher information as they do for the binomial; the intention in using them is to avoid the anomalous behaviour of the most conservative prior for extremely small sample sizes (Section 2.8).

As with the beta Jeffreys method, stratified sampling is handled by drawing from posteriors on each stratum. As the number of strata increases, the size of each stratum decreases, and the risk that the prior dominates the evidence increases. In addition, stratification separates heterogeneous sub-populations, making unbalanced populations more likely, and again raising the risk of a dominating prior. For both these reasons, stratified sampling emphasises the importance of careful prior choice.

Also in common with the Jeffreys method, even if there are no relevant documents in the unretrieved sample, the upper limit of the recall interval will not be strictly one, though rounding effects may make it effectively one. For very small samples, the gap from one may be noticeable. As with the Jeffreys method, this gap can be addressed by setting the upper limit to one if there are no relevant documents in the unretrieved sample, but that is not done here.

4. EVALUATION

Our experiments test the interval estimators described in Section 3 for their coverage of three representative retrieval estimation scenarios (Section 4.1). The coverage performance of each method has been previewed in Table IV; the beta-binomial posterior with most conservative prior gives the most accurate, most balanced, and most stable coverage of the nine methods considered (Section 4.2). We offer advice on sampling design, the allocation of assessments to retrieved and unretrieved segments, and the interval widths that can be expected (Section 4.3). Finally, we calculate beta-binomial

MCP intervals for the participants in the Interactive Task of the Legal Track in TREC 2008 and TREC 2009 (Section 4.4).

4.1. Evaluation methodology

The coverage that an interval gives a certain parameter value is the proportion of samples for which that value falls inside the interval; mean coverage is the average across a given distribution of parameter values (Section 2.5). We propose three criteria for evaluating the coverage of an approximate confidence interval:

Unbiasedness. Is mean coverage across different values of the population parameter at the nominal level?

Consistency. Is the variability in coverage between different parameters small?

Balance. Is the proportion of coverage misses the same at the lower end as the upper?

The term “unbiased” is sometimes used in reference to confidence intervals as meaning that the interval is at least as likely to cover the true parameter value as it is to cover any other [Guenther 1971]; here, we instead use it in the broader statistical sense that the average value (across our sample space of parameter values) is close to the predicted one.⁷ Unbiasedness in the former, more technical sense is secondary to accuracy of coverage, and is difficult to assess empirically.

Another criteria proposed in the literature for evaluating confidence intervals is narrowness in parameter space, with the narrower intervals preferred to wider ones if they have the same accuracy of coverage [Smithson 2002]. Narrowness is secondary to exactness of coverage, and is tricky to interpret for approximate intervals; intervals providing under-coverage will tend to (though not always) be shorter than exact ones, without this being a positive feature of the interval. In addition, it is not clear that all ranges of parameters have the same value; we may care more, for instance, about fidelity at the extremes than towards the middle. Nevertheless, interval width provides an important additional constraint. A method that gave an interval of $[0, 1]$ 95% of the time, of $(0, 0)$ 2.5% of the time, and of $(1, 1)$ 2.5% of the time, would achieve perfect performance under the criteria of unbiasedness, consistency, and balance, and yet convey no useful information about the true parameter.⁸ A consideration of interval width, though, would reveal the interval’s degeneracy. We therefore also examine mean interval width for each scenario and method, as a sanity check on interval behaviour.

Coverage performance depends upon the distribution over the population parameter. In Section 2.6, binomial intervals were evaluated for a uniform distribution over prevalence. But a uniform distribution over segment population and yield in the retrieval case makes little sense. Instead, we evaluate recall intervals against different retrieval scenarios. Each scenario defines distributions over population parameters, such as the size of the population, the sample size, and the effectiveness of the retrieval. Observed values are drawn for each parameter to instantiate a realization, and samples are repeatedly drawn from a realization to calculate coverage of that realization. A set of such realizations are drawn, and the coverage characteristics of the scenario observed from the set.

Table V lists the parameters for which each scenario defines distributions. All values described in Table I can be calculated from these parameters, except for the r_0 and r_1 , the number of relevant documents in the sample from each segment. For instance, N_1 , the retrieval segment size, is $\lfloor N \cdot \text{Rec}/\text{Prec} \rfloor$ ($\lfloor \cdot \rfloor$ is the rounding operator), while R_1 ,

⁷We are, quite frankly, out of suitable synonyms, “accurate”, “precise”, and “exact” also having other meanings when referring to confidence intervals.

⁸I owe this observation to Dave Lewis.

Table V. Scenario variables.

Variable	Meaning
N_*	Size of population
$\pi_* = R_*/N_*$	Proportion relevant in population
$\text{Rec} = R_1/R_*$	Recall of retrieval
$\text{Prec} = R_1/N_1$	Precision of retrieval
n_1	Size of sample from retrieved segment
n_0	Size of sample from unretrieved segment

Table VI. Neutral scenario.

Variable	Range	Mean	Distribution
N_*	1,000 to 4,100,000	495,000	$1000 \cdot 2^{Unif(0,12)}$
π_*	0.02 to 0.72	0.287	$0.02 \cdot Unif(1, 6)^2$
Rec	0.1 to 1.0	0.55	$Unif(0.1, 1.0)$
Prec	0.1 to 1.0	0.64	$Unif(\min(0.1, 0.95\pi_*, 1.05R_1/N_*), 1.0)$
n_1	10 to 10,000	890	$10 \cdot 2^{Unif(0, \max(10, \lfloor \log_2(N_1/10) \rfloor))}$
n_0	10 to 10,000	1,150	$10 \cdot 2^{Unif(0, \max(10, \lfloor \log_2(N_0/10) \rfloor))}$

Table VII. Legal scenario.

Variable	Range	Mean	Distribution
N_*	500,000 to 50,000,000	1,075,000	$5 \times 10^5 \cdot 10^{Unif(0,2)}$
π_*	0.003 to 0.115	0.031	$2 \times 10^{-3} \cdot 1.5^{Unif(1,10)}$
Rec	0.0025 to 0.84	0.33	$2.5 \times 10^{-3} \cdot Unif(1, 34)^{1.65}$
Prec	0.025 to 0.92	0.48	$Unif(\min(0.025, 2 \cdot R_1/N_*), 0.92)$
n_1	20 to 5,120	820	$20 \cdot 2^{Unif(0, \max(8, \lfloor \log_2(N_1/20) \rfloor))}$
n_0	100 to 12,800	3,170	$100 \cdot 2^{Unif(0, \max(7, \lfloor \log_2(N_0/100) \rfloor))}$

the number of relevant documents in the retrieved segment, is $\lfloor N \cdot \text{Prec} \rfloor$. The number of relevant documents in the retrieved segment, r_1 , is observed for a realized scenario by drawing a sample of size n_1 from the segment population (which is equivalent to taking an observation from a $HyperGeom(N_1, R_1, n_1)$ distribution). The number of relevant documents in the unretrieved segment, r_0 , is similarly observed. We define three scenarios using these variables, described next.

4.1.1. Neutral scenario. The neutral scenario covers a range of retrieval and classification tasks, on populations large enough to require sampling. Table VI describes the range and means of the scenario variables in the neutral scenario, and the underlying distributions assigned to these variables. (Means are analytic or large-sample estimates of distribution means, not those observed on the actual sample drawn for the experiments.) Two lower bounds are set on precision, the first to ensure precision of the retrieval segment is not much worse than of the unretrieved (the retrieval system does at least as well as random), and the second to guarantee that not all documents are retrieved.

4.1.2. Legal scenario. The legal scenario (Table VII) represents e-discovery retrieval, as reflected in the Interactive Task of the TREC Legal Track (Section 4.4), and by extension other large-scale retrievals. Scenario parameter distributions are fitted to those observed amongst TREC topics and participants in 2008 and 2009. For instance, Figure 9 shows the fit of the recall distribution in Table VII to the recall values observed in the legal track. The population of documents in the scenario is large, the proportion relevant small (though not as small as in some other information retrieval

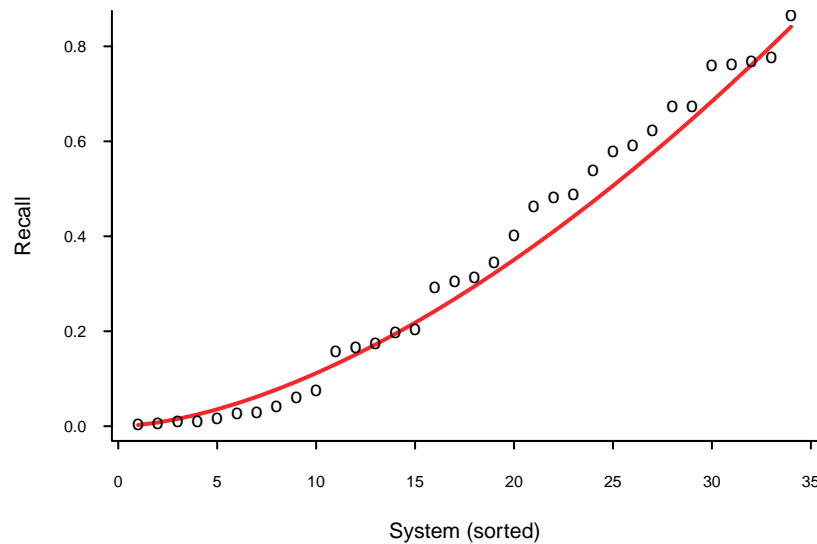


Fig. 9. Sorted recall of participants in the Interactive Task of the TREC 2008 and TREC 2009 Legal Track (dots), with fitted line of $2.5 \times 10^{-3} \cdot \text{Unif}(1, 34)^{1.65}$. Recall scores are sample-based estimates using adjudicated assessments [Oard et al. 2008].

Table VIII. Small scenario.

Variable	Range	Mean	Distribution
N_*	1,000 to 10,000	3,900	$1000 \cdot 10 \text{Unif}(0,1)$
π_*	0.02 to 0.23	0.085	$0.02 \cdot 1.5 \text{Unif}(0,6)$
Rec	0.1 to 1.0	0.55	$\text{Unif}(0.1, 1.0)$
Prec	0.025 to 0.92	0.51	$\text{Unif}(\min(0.025, 2R_1/N_*), 0.92)$
n_1	20% to 50% of retrieved segment	160	$N_1 \cdot \text{Unif}(0.2, 0.5)$
n_0	5% to 30% of unretrieved segment	600	$N_0 \cdot \text{Unif}(0.05, 0.3)$

tasks), and the sample budget is constrained. As a result, the unretrieved sample will often contain few, or even no, relevant documents, testing interval estimators' handling of extreme proportions. The lower bound on precision ensures that no more than half the corpus is retrieved. Correlation between precision and recall amongst the TREC participants is a statistically non-significant $\rho = 0.25$, the only relationship being that very high recall is not matched with very low precision; the scenario lower bound on precision effectively captures this weak relationship. Sample sizes are larger absolutely than for the neutral scenario, but a smaller proportion of the population. The median retrieved sample takes in 2% of the retrieved segment, while the median unretrieved sample takes in just 0.5% of its segment. All of these values agree with those observed in the legal track.

4.1.3. Small scenario. The small scenario, described in Table VIII, shrinks population size to test how well the intervals handle the finite population case. Sample sizes are made a substantial proportion of the retrieved and unretrieved populations, up to 50% in the former case, and up to 30% in the latter. Prevalence is kept below 0.25, in line with most retrieval tasks (in contrast, binary classification tasks can readily have prevalences at or above 0.5). The retrieval is constrained to produce no more than half the population.

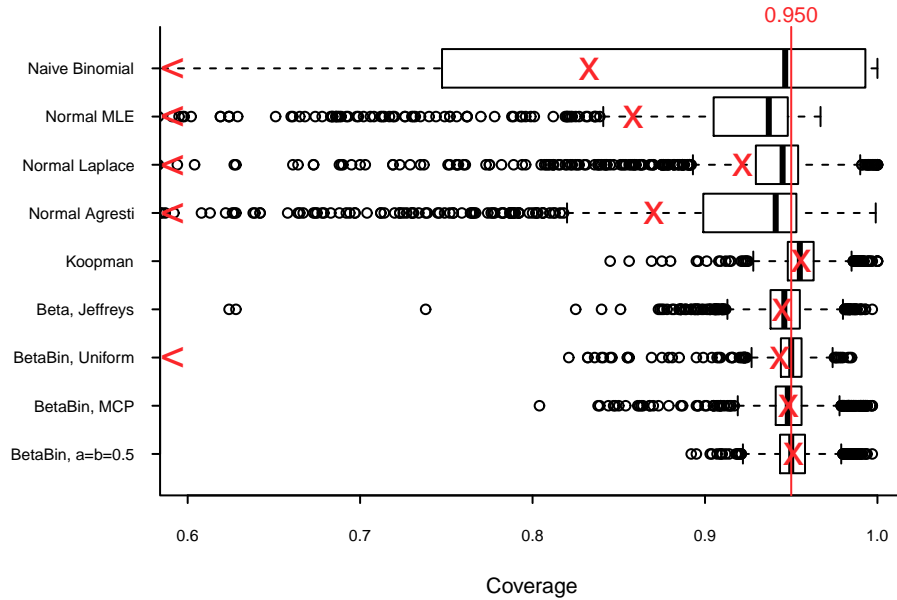


Fig. 10. Coverage of the different recall interval estimators for the neutral scenario. A total of 1,000 scenario realizations are instantiated, and 1,000 samples drawn for each realization. The same realizations and samples are used for each of the interval methods. The coverage of a realization is the proportion of samples from the realization producing a confidence interval that covers the realization's true population recall. The thick line in each box shows median coverage across these realizations, the box edges the first and third quartiles. Circles show realizations with coverage outside the quartiles by more than 1.5 times the interquartile range. Mean coverage is shown as a cross. Methods for which minimum coverage is below the range of the graph are marked on the left margin with "<".

4.2. Coverage of the three scenarios

We now report the coverage of the nine different confidence interval estimators on the three scenarios described above, testing the three criteria established at the start of Section 4: unbiasedness (mean coverage near the nominal level of 95%); consistency (little variability in coverage); and (in Section 4.2.4) balance (parameter as likely to be missed below the interval as above).

4.2.1. Neutral scenario. Figure 10 shows the mean and median coverage (cross and bar) and coverage consistency (width of box and whiskers and location of outliers) of the neutral scenario by the different interval methods. The naive binomial is highly inconsistent, since its assumptions are grossly violated by different sampling rates between retrieved and unretrieved segments (Section 3.1). The MLE normal is unstable and biased towards under-coverage (that is, the postulated interval is too narrow), with mean coverage around 0.85; the cause is presumably understatement of variance (Section 3.2). The Laplace adjustment largely corrects the bias, but still has instances of significant under-coverage, while the Agresti-Coull adjustment is little better than the MLE method, suggesting the plus-two adjustment is too large. The remaining five methods provide better consistency, but the Koopman ratio of binomials method is biased high (interval too wide), and the Jeffreys ratio of betas slightly low; these tendencies are even stronger in the small scenario (Section 4.2.3). The uniform beta-binomial posterior method is also biased low. The most conservative prior is unbiased, but also shows a tendency to noticeable under-coverage. The best results are shown by the beta-

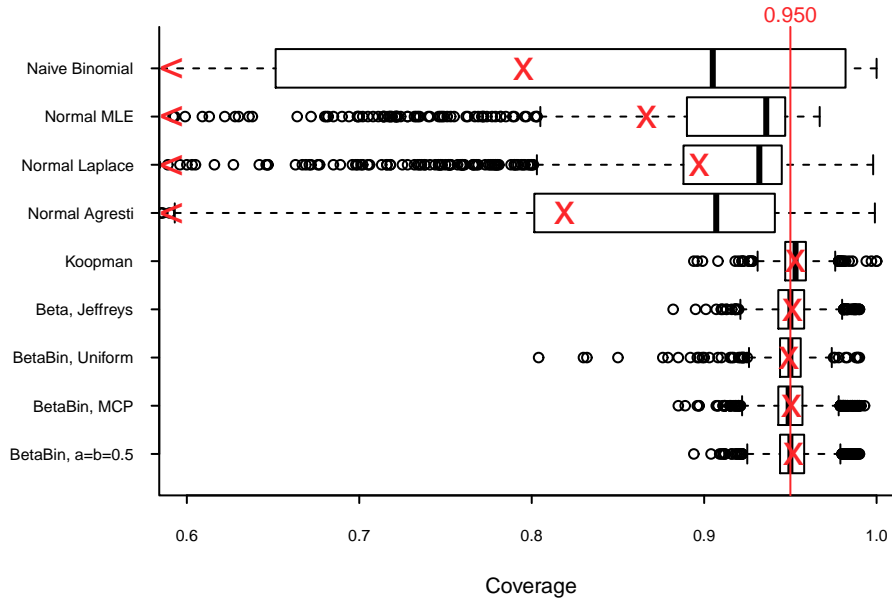


Fig. 11. Coverage of the different confidence interval estimators for the legal scenario. Other details are as for Figure 12.

binomial with the simple prior of $\alpha = \beta = 0.5$. Choice of prior matters; but the most satisfactory prior is not the theoretically most grounded one.

4.2.2. Legal scenario. The legal scenario tests the ability of the intervals to cope with extreme sample prevalences; in particular, the likelihood that the unretrieved sample will have few or no positive instances in it. The results are shown in Figure 11. The naive binomial and normal methods perform even worse than for the neutral scenario. The Laplace adjustment to the normal provides only partial correction for the extreme prevalence problem, while the Agresti-Coull adjustment actually makes under-coverage worse. Analysis shows that almost all under-coverage in the adjusted methods is due to understating recall, by overstating unretrieved yield. In short, the adjustment is too large, especially the plus-two for Agresti-Coull. The remaining five methods all give accurate, consistent coverage. The Koopman method is most consistent, though biased slightly high. The uniform prior to the beta-binomial ratio gives occasional marked undercoverage; again, the implicit plus-one adjustment is in some cases too high. The most conservative prior avoids this problem, as does using the hyperparameters $\alpha = \beta = 0.5$.

4.2.3. Small scenario. The small scenario tests the finite-population case of sample size being a significant proportion of population size. The results are given in Figure 12. The naive binomial and the three normal methods again show bias and poor consistency; the Laplace-adjusted normal is the best of them. Unlike the previous scenarios, however, the Koopman and Beta Jeffreys methods are strongly biased. The Koopman method ignores the reduction in uncertainty caused by having observed a large fraction of the population, and produces intervals that are too wide. Conversely, the binomial-based beta posterior misses the dependence between sampled and unsampled prevalence (a positive in the sample is one less positive in the remainder), and gives intervals that are too narrow. The beta-binomial posterior methods are largely accurate,

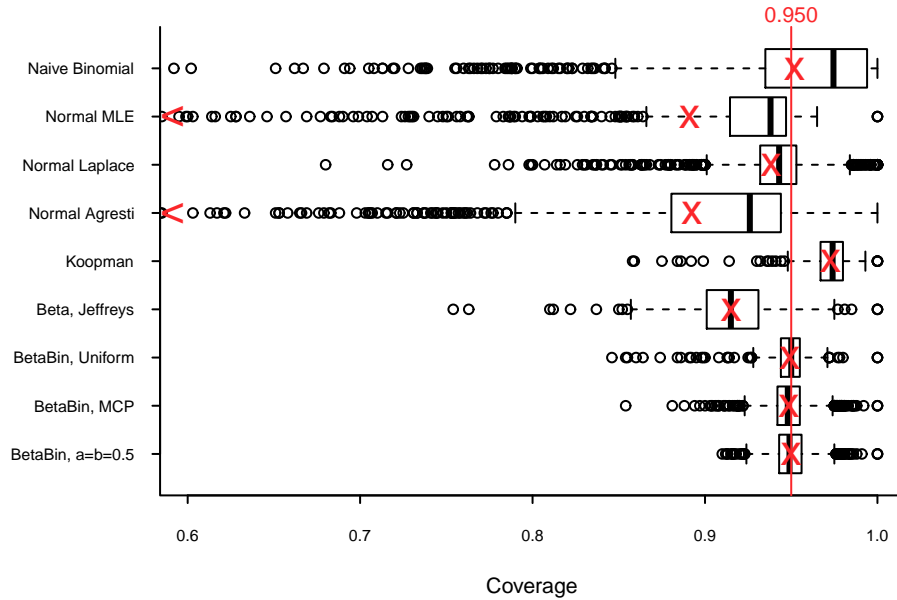


Fig. 12. Coverage of the different confidence interval estimators for the small scenario. Other details are as for Figure 12.

though as with the neutral scenario, the uniform and most-conservative priors show a tendency to marked under-coverage that the $\alpha = \beta = 0.5$ prior avoids.

4.2.4. Upper and lower gaps. The third desideratum of an interval, after unbiasedness and consistency, is that it be balanced, with missed parameters as likely to fall above as below the interval. Interval balance for the neutral scenario is shown in Figure 13. The MLE normal method tends to overstate recall (true recall more likely to fall below the interval than above it), and the adjusted normal methods to understate it, because they respectively understate and overstate prevalence in the unretrieved segment. The uniform prior to the beta-binomial posterior tends to understate recall, even though its coverage is unbiased (Figure 12), again because its implicit adjustment to the unretrieved segment is in some circumstances too large. The most conservative prior, in contrast, tends to understate recall, while the prior with $\alpha = \beta = 0.5$ is relatively balanced. The legal and small scenarios (not shown for space) give similar results for balance, with the $\alpha = \beta = 0.5$ beta-binomial being consistently the best balanced.

4.2.5. Discussion. We summarize the bias and consistency characteristics of the examined interval methods in Table IX, as the root mean squared error (RMSE) between actual and nominal coverage. The naive binomial method is biased and highly unstable for all the scenarios considered, due to the unequal sampling between retrieved and unretrieved segments. The normal methods are less biased, with the Laplace adjustment offering reasonable median performance, but are highly unstable, with coverage often falling far below the nominal level. These interval estimators should be avoided. The Koopman ratio-of-binomial method, and the beta and beta-binomial posterior methods, offer more reliable performance. For the neutral and legal scenarios, they are equally unbiased and stable. For the finite population case of the small scenario, however, the Koopman method leads to over-coverage, and the Jeffreys to under-coverage, each due to its particular failure to handle the limited population case; the

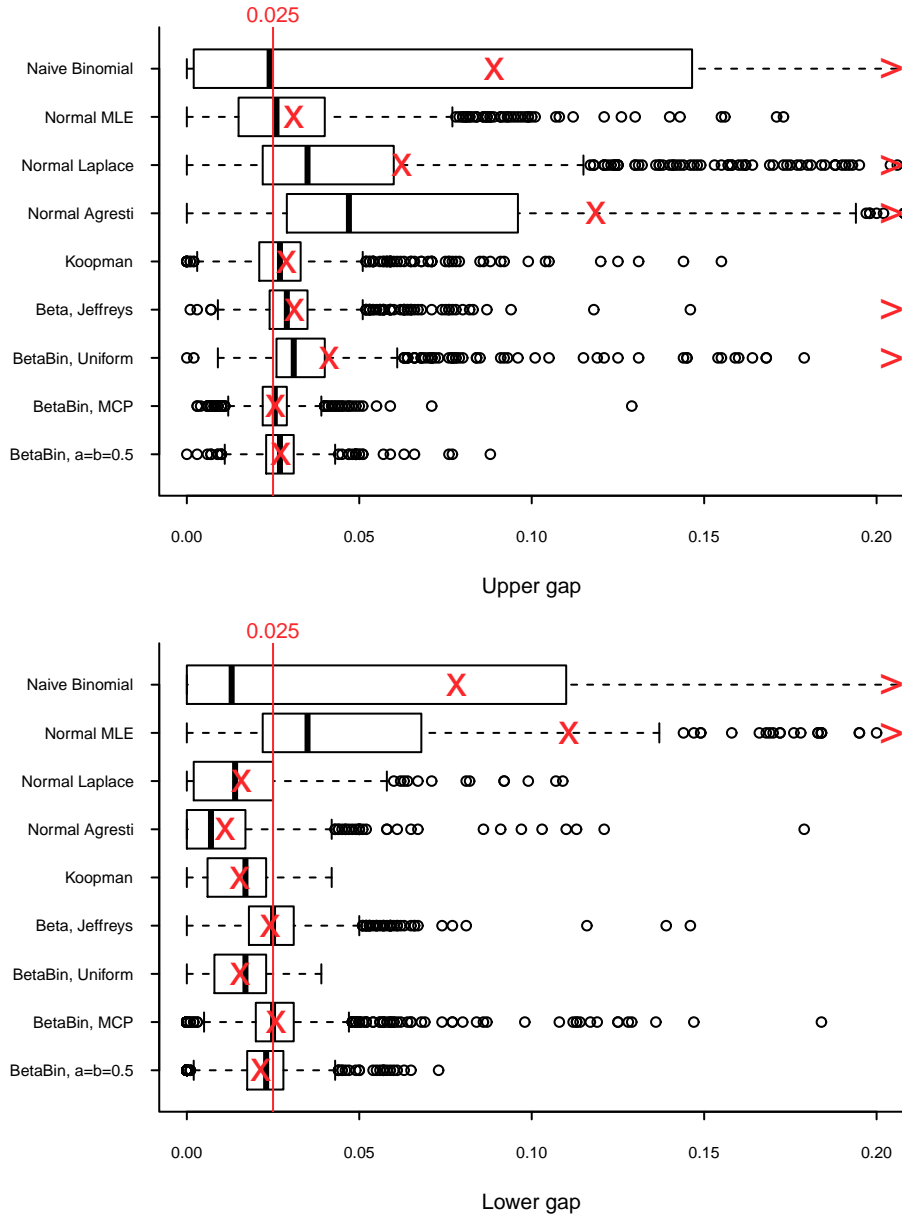


Fig. 13. Upper and lower gaps for different CI methods on the neutral scenario. Each underlying data point represents a scenario realization, and states the proportion of the samples drawn from that realization for which recall falls above (or below) the interval estimated from the sample. Realizations and resamples are as for Figure 12. The thick lines shows the median gap, the cross the mean. Box edges are first and third quartiles, while circles show gaps that are outside the quartiles by more than 1.5 times the inter-quartile range. Methods for which the maximum gap is above the range of the graph are marked on the right margin with “>”.

Table IX. Root mean squared error from nominal coverage for interval estimators on the three different scenarios.

Method	Scenario			Mean
	Neutral	Legal	Small	
Naive Binomial	0.252	0.281	0.063	0.199
Normal MLE	0.225	0.189	0.155	0.189
Normal Laplace	0.107	0.110	0.034	0.084
Normal Agresti	0.215	0.251	0.120	0.195
Koopman	0.016	0.011	0.026	0.017
Beta, Jeffreys	0.026	0.014	0.042	0.027
BetaBin, Uniform	0.072	0.014	0.012	0.033
BetaBin, MCP	0.020	0.015	0.014	0.016
BetaBin, a=b=0.5	0.014	0.013	0.012	0.013

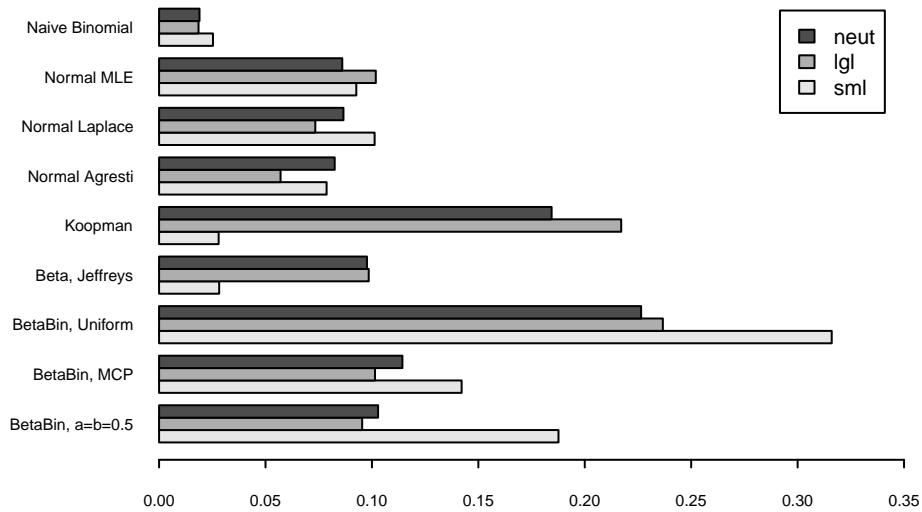


Fig. 14. Proportion of scenario realizations for which each interval method gives coverage closest to the nominal level, for each of the evaluation scenarios. Tied coverage (to the fidelity of the simulation, which is 1,000 simulated samples per realization) receives split counts.

conservatism of the Koopman method may make it preferable between the two. The neutral and most conservative priors to the beta-binomial method offer similar unbiasedness and consistency on the neutral and small scenarios, but the uniform prior tends more to undercoverage on the legal scenario. The simple $\alpha = \beta = 0.5$ prior to the beta-binomial gives superior coverage characteristics compared to the other beta-binomial methods, as well as the other method families considered here; it also offers the best balance (as shown in Figure 13). For the above reasons, the beta-binomial posterior interval with hyperparameters α and β set to 0.5 is the preferred method.

The RMSE results in Table IX allow us to compare the overall performance of the different interval methods. But is it the case that methods that perform better overall can be relied upon to perform better in all or most cases? Figure 14 shows that the answer is no. Though the Koopman and posterior methods have as little as a tenth of the RMSE than the normal methods on the uniform scenario (first column of Table IX),

Table X. Mean width of 95% recall interval for different interval methods on the three scenarios, along with the mean coverage of these methods.

Method	Neutral		Legal		Small	
	width	cvr	width	cvr	width	cvr
Naive Binomial	0.23	0.83	0.21	0.80	0.29	0.95
Normal MLE	0.18	0.86	0.21	0.87	0.21	0.89
Normal Laplace	0.22	0.92	0.23	0.90	0.22	0.94
Normal Agresti	0.19	0.87	0.19	0.82	0.21	0.89
Koopman	0.22	0.96	0.26	0.95	0.24	0.97
Beta, Jeffreys	0.22	0.94	0.26	0.95	0.19	0.92
BetaBin, Uniform	0.21	0.94	0.24	0.95	0.22	0.95
BetaBin, MCP	0.22	0.95	0.26	0.95	0.21	0.95
BetaBin, a=b=0.5	0.22	0.95	0.26	0.95	0.21	0.95

they outperform only on a moderate majority of individual realizations. As strikingly, although the most-conservative prior and $\alpha = \beta = 0.5$ prior beta-binomial both have half the RMSE of the uniform prior, it is the uniform prior that comes closer to the nominal coverage in the majority of realizations for all three scenarios. The weaker methods underperform not by being slightly less accurate on the majority of cases, but by being much less accurate in particular circumstances. The evaluator, however, cannot be sure from the sample alone which true circumstance the population fits into; therefore, selection of the overall most accurate interval remains advisable.

The mean interval width of the different methods is given in Table X; for ease of reference, the mean coverage of these methods is also provided (previously shown in Figures 10, 11, and 12). Mean interval width, at around 0.22, is similar between different scenarios, though standard deviation (not shown) is lower for the small (median 0.12) than for the legal (0.19) and neutral (0.20) scenarios. Mean width does not always co-vary with mean coverage; for instance, the naive binomial method gives the lowest mean coverage but the widest mean intervals for the neutral scenario. The methods with good coverage performance, though, have similar mean interval widths; we can choose between them based on their coverage without concern that degenerate behaviour in width will result.

4.3. Sampling design and expected intervals

The previous sections have discussed the retrospective analysis of recall confidence intervals. The experimental designer or system auditor, however, is faced initially with several prospective questions. How wide a confidence interval is an analysis likely to produce? How large a sample size is needed to bring this interval within reasonable bounds? And how should sampled assessments be allocated between the retrieved and unretrieved segments to minimize the interval?

Interval width depends on sample size; division of sample amongst segments; corpus yield; recall; retrieval size; and (though only marginally) population size. Sample size and sample allocation are under the evaluator's control, though the former will be limited by cost, and the latter may be constrained by other evaluation goals; retrieval and population size are known at sample time; but yield and recall are unknown prior to sampling. Figure 15 shows the relationship between sample allocation and interval width for varying retrieval sizes, given a retrieval process with estimated recall and precision of 0.5. For higher corpus yield, the recall interval is narrower, and not too sensitive to sample allocation. As yield decreases, however, the interval widens. Moreover, lower prevalence in the unretrieved segment increases the variance of unretrieved yield estimates, and more assessments must be allocated to the segment to compensate. The anomalous fall in the interval width for the 5,000-size retrieval as

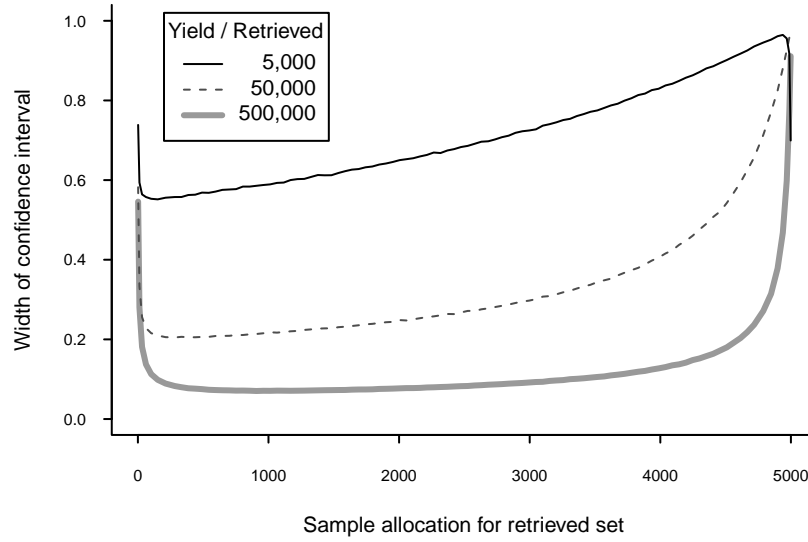


Fig. 15. The width of confidence intervals, using the beta-binomial posterior with $\alpha = \beta = 0.5$, on sampling n documents from the retrieved segment and $5,000 - n$ documents from the unretrieved segment, for a population of 5,000,000 documents, and different corpus yields R_* . The recall and the precision of the retrieval are both 0.5, making retrieval size equal to yield. Proportions relevant in the sample are assumed to be the same as proportions relevant in the population; where this would result in fractional numbers of positive in the sample, we abuse our interval estimator by feeding these fractional values in.

the sample is almost entirely allocated to the retrieved segment, and the retrieved segment is almost fully sampled, is an artifact of the behaviour of the posterior mentioned at the end of Section 3.6: namely, that the upper limit of the interval never goes to 1, and that with extremely small sample sizes it can fall well short. This anomaly can be addressed by setting the upper limit on recall to 1 if there are no relevant documents in the unretrieved sample.

How sensitive is the optimal allocation to actual retrieval performance? Figure 16 examines this question. Retrieval quality clearly affects interval width. High precision and low recall leaves high prevalence in both retrieved and unretrieved segments, reducing variance of the yield estimate and so interval width. Moreover, in this circumstance, width is not sensitive to allocation. Low precision and high recall have the opposite effect, causing wide intervals and pushing assessments towards the unretrieved set. A 20% allocation towards the retrieved stratum is not too far from optimal for any of the scenarios here, but may conflict with the accurate estimation of precision, or a desire to thoroughly survey the retrieved documents. Allocating more assessments to the retrieved set, however, risks widening the recall interval substantially, depending up on the retrieval's actual effectiveness.

Allocation aside, the chief variable at the evaluator's control for minimizing interval width is sample size. Figure 17 shows interval width as a function of sample size for the beta-binomial with $\alpha = \beta = 0.5$ and MLE normal methods, assuming an ideal sample allocation between segments. The normal method shows the standard $1/\sqrt{n}$ reduction in interval width with an n -fold increase in sample size, but gives overly wide intervals for low prevalence and small samples, producing intervals wider than 1.0 in some cases. The beta-binomial agrees closely with the normal for high prevalence and

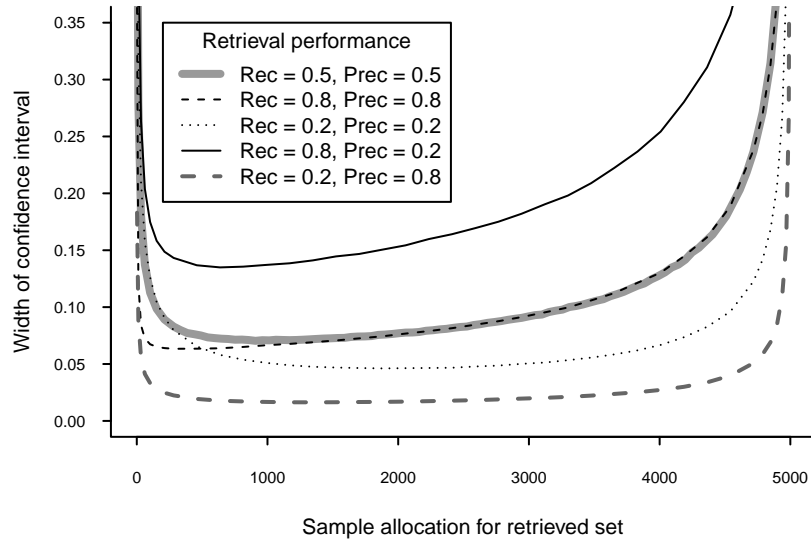


Fig. 16. The width of confidence intervals, using the beta-binomial posterior with $\alpha = \beta = 0.5$, for retrievals of different levels of effectiveness. A sample of n is drawn from the retrieved segment and $5,000 - n$ documents from the unretrieved segment, from a population of 5,000,000 documents, and a retrieved segment size of 500,000.

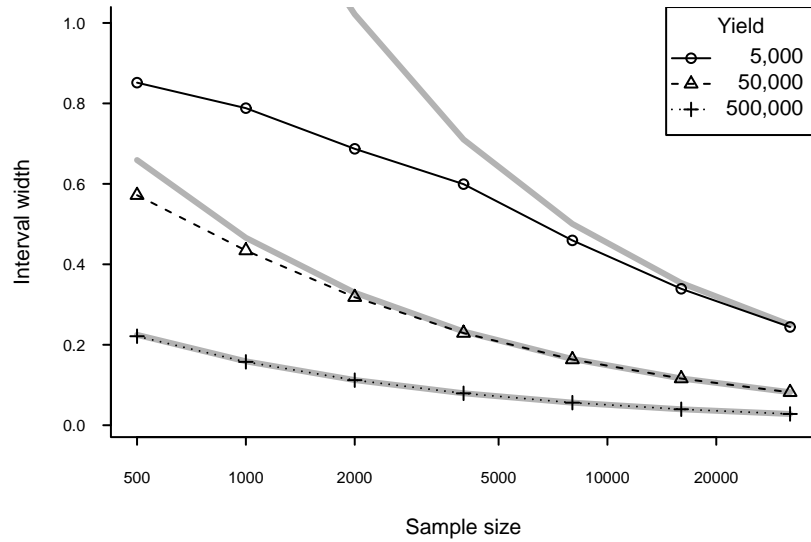


Fig. 17. Width of recall confidence interval for varying sample sizes and different retrieved segment sizes. The population size is 5×10^6 , and the retrieval has recall and precision both 0.5. Intervals are shown for the optimal allocation of samples and assessments to the retrieved and unretrieved segments. The black lines show the beta-binomial posterior, with $\alpha = \beta = 0.5$. The grey lines show the intervals given by MLE normal method.

Table XI. Recall estimates with 95% confidence intervals for the Interactive Task of the TREC 2008 Legal Track. The beta-binomial interval uses the $\alpha = \beta = 0.5$ as prior. The normal interval use MLE variance estimates, without (“Naive”) and with (“Corr”) correction for the correlation between system and corpus yield estimates.

Topic	Team	Recall	CI (BBin, 0.5)		CI (Norm, naive)		CI (Norm, corr.)	
			2.5%	97.5%	2.5%	97.5%	2.5%	97.5%
t102	Ad Hoc Pool	0.314	0.273	0.360	0.266	0.362	0.271	0.358
	Clearwell	0.016	0.014	0.018	0.014	0.018	0.014	0.018
	Pittsburgh	0.007	0.006	0.008	0.006	0.008	0.006	0.008
t103	H5	0.624	0.580	0.665	0.579	0.668	0.581	0.666
	Ad Hoc Pool	0.403	0.374	0.430	0.371	0.434	0.374	0.431
	Clearwell	0.158	0.146	0.169	0.146	0.169	0.146	0.169
	Buffalo	0.061	0.056	0.066	0.056	0.066	0.056	0.066
	Pittsburgh	0.026	0.024	0.028	0.024	0.029	0.024	0.029
t104	Ad Hoc Pool	0.345	0.185	0.578	0.111	0.580	0.143	0.548
	Clearwell	0.003	0.002	0.005	0.001	0.004	0.001	0.004

large samples, but gives smaller intervals for low prevalence and small samples. Consequently, the decrease in interval width is slower than $1/\sqrt{n}$ in such circumstances.

4.4. Intervals for TREC Legal Track

We conclude this experimental section by calculating confidence intervals for the Interactive Task of the TREC 2008 and TREC 2009 Legal Track. Task evaluation employed stratified sampling (Section 2.2), with strata defined by the intersection of participant team retrievals; n teams define 2^n strata, some of which may be empty. Different strata received different sampling rates; in particular, the bottom stratum, of documents retrieved by no system, was sampled only sparsely [Oard et al. 2008]. The recall confidence interval for the official results was calculated using the Normal-MLE method (Section 3.2), though with a propagation of error expression that omits yield covariance (Section 2.10).

Table XI compares the normal and beta-binomial confidence intervals for TREC 2008. The “Ad Hoc Pool” team was a deep pseudo-run made up by pooling the automated runs from a parallel task. The normal intervals use the MLE estimate of variance, and are calculated without and with correction for the correlation between team and corpus yield estimates (Section 2.10); the naive or uncorrected interval estimate is as reported in the official results [Oard et al. 2008]. The correlation-corrected interval differs substantially from the naive only for the Ad Hoc Pool pseudo-team on Topic 104, since only then does it display the necessary combination of a high-recall, low-precision (0.023) run. Also only on Topic 104 is there a substantial difference between the intervals of the corrected MLE normal and the beta-binomial methods, since the unretrieved segment sample prevalence is not small for the other two topics.

The normal and beta-binomial intervals for TREC 2009 are shown in Table XII. Correlation correction is applied for the normal version. The uncorrected intervals (not shown) are marginally narrower in most cases, and sharply narrower for a few. For instance, the interval for UW on Topic 203 (once again, a retrieval with high recall but moderate precision) is $[0.765, 0.964]$ without correlation correction, $[0.832, 0.897]$ with. In contrast to the previous year, the normal and beta-binomial intervals generally disagree. The mean absolute difference in width is 13%, with the beta-binomial being narrower for all topics save Topic 203. The beta-binomial generally gives a lesser lower bound, and is wider beneath than above the point estimate. The cause is low or zero sample prevalence in the bottom stratum, to which the beta-binomial method correctly gives a wider upper than lower bound. The normal method also assigns an impossible

Table XII. Recall and interval estimates for the Interactive Task of the TREC 2009 Legal Track. The MLE normal and Laplace methods correct for yield estimate correlation.

Topic	Team	Recall	CI (BBin, 0.5)		CI (Norm, corr.)		CI (Laplace)	
			2.5%	97.5%	2.5%	97.5%	2.5%	97.5%
t201	UW	0.778	0.451	0.929	0.485	1.070	0.351	0.916
	CS	0.489	0.287	0.593	0.304	0.673	0.226	0.593
	CB	0.204	0.119	0.250	0.126	0.282	0.093	0.247
	UP	0.167	0.097	0.208	0.102	0.231	0.076	0.205
t202	UW	0.673	0.523	0.768	0.542	0.803	0.501	0.771
	CS	0.579	0.451	0.662	0.467	0.692	0.432	0.664
t203	UW	0.865	0.625	0.881	0.832	0.897	0.552	0.940
	UB	0.592	0.427	0.621	0.552	0.631	0.378	0.652
	ZL-NoCull	0.175	0.123	0.185	0.157	0.192	0.105	0.185
	ZL-Cull	0.029	0.019	0.034	0.023	0.035	0.016	0.033
t204	H5	0.762	0.572	0.882	0.595	0.930	0.538	0.874
	UD	0.305	0.229	0.360	0.235	0.374	0.216	0.358
	CB	0.198	0.148	0.240	0.151	0.246	0.140	0.238
t205	CS	0.673	0.593	0.738	0.599	0.747	0.587	0.737
	EQ	0.463	0.404	0.509	0.409	0.516	0.399	0.505
	IN	0.292	0.253	0.327	0.254	0.329	0.250	0.325
t206	CB-High	0.076	0.052	0.112	0.046	0.106	0.049	0.109
	LO	0.042	0.025	0.069	0.020	0.063	0.023	0.067
	CB-Mid	0.011	0.007	0.015	0.007	0.015	0.007	0.014
	CB-Low	0.009	0.006	0.013	0.006	0.013	0.005	0.012
t207	CB	0.768	0.698	0.798	0.723	0.813	0.689	0.802
	UW	0.761	0.689	0.791	0.714	0.807	0.677	0.791
	LO	0.538	0.489	0.566	0.503	0.573	0.484	0.569
	EQ	0.483	0.437	0.507	0.451	0.515	0.430	0.507

upper-bound recall of 1.07, even though the team fails to return several known relevant documents.

For comparison, the final two columns of Table XII show the effect on the normal interval of the Laplace adjustment (Section 3.3). The adjusted interval is generally left-skewed, like the beta-binomial, through its handling of low-prevalence unretrieved strata. The intervals, however, are wider, sometimes substantially so, than the beta-binomial. Adding the plus-one adjustment to every stratum is excessive. A smaller adjustment might be considered, but the most conservative prior to the beta-binomial offers a more principled approach.

5. CONCLUSION

Several methods for calculating a confidence interval on a recall estimate have been considered in this article. The most straightforward method takes maximum-likelihood estimates of the sampling variance of yield on each segment, aggregates them to a sampling error on recall, and derives bounds via a normal approximation. While this approach provides the nominal mean coverage in some circumstances, it drastically over- or under-covers true recall in others. A particular weakness of the method is that it understates variance for extreme prevalences. This weakness can be addressed using adjustments, such as the Laplace plus-one and the Agresti-Coull plus-two approaches. But while these adjustments give appropriate corrections in some scenarios, they over-correct in others. They do not address other objections, such as the non-normal sampling distribution of recall; and they generalize poorly to the stratified case.

Alternative methods to the normal approximation have been considered. The naive binomial method, used for bounding sensitivity in equal-sampling setups, performs as poorly as the violation of its assumptions would lead one to expect. The Koopman analytical ratio-of-binomial estimator gives better results, as does the ratio-of-binomials posterior to the Jeffreys prior. Both, however, are inaccurate in the finite population case.

The most satisfactory approximate interval defines a beta-binomial posterior for the retrieved and unretrieved segments, and samples from this posterior to generate a distribution over recall. This method gives the most accurate coverage in all of the scenarios considered here. It is unbiased, with mean coverage sitting on the nominal confidence level, and it is comparatively stable, with quartiles clustering to within a couple of percentage points of the nominal level. Choice of prior matters, but the best prior is not necessarily the most theoretically grounded one. Both the simple uniform prior and the theoretically-based most conservative prior suffer from occasional marked under-coverage. Best performance is given by setting the hyperparameters α and β to 0.5.

We have provided guidance on the width of confidence intervals that can be expected in practice, and an experimental design to reduce this width. In typical retrieval evaluation scenarios, the predominant factor outside the experimenter's direct control that influences interval width is prevalence in the unretrieved segment; the lower this prevalence is, the wider the confidence interval on recall. For low prevalence values, the interval is minimized by devoting the majority of assessment resources to the unretrieved segment; but such an allocation mitigates against accurate precision estimates and other practical objectives, such as characterizing the relevant set. Assuming optimal allocation between segments, interval width under the beta-binomial prior method decreases by the usual square root of sample size only once the interval is well below 0.5; before that, decrease in interval width is slower.

Finally, we have used most-conservative beta-binomial posterior method recommended in this article to produce interval estimates on the TREC 2008 and TREC 2009 Legal Track, Interactive Task participants, and compared them with the naive, unadjusted, and adjusted normal methods. In some cases, the normal methods give similar estimates to the beta-binomial; in others, marked differences exist, particularly where sample prevalence in the unretrieved segment is low.

5.1. Future work

The methods considered in this article produce approximate confidence intervals, not exact ones. "Approximate" is not a disparaging term here, nor "exact" a laudatory one. Rather, the goals of the two methods are different: approximate methods seek mean coverage at the nominal level; exact ones ensure nominal coverage in the worst case, at the expense of over-coverage on average. Coverage of the beta-binomial rarely falls far below the nominal level. In some circumstances, however, such as certification before a court, a guarantee of minimum coverage may be required. For such cases, an exact method is needed.

Many factors affect the width of the recall interval. Guidance has been provided to the evaluation designer on some of them, including the allocation of samples to the retrieved and unretrieved segments, and the reduction in width with sample size. These decisions rely in part, though, on the retrieved and corpus yields, which are not known prior to sampling. Distributions might be posited over these parameters, perhaps based on past experience or evaluator speculation, and an allocation chosen that minimizes in expectation some risk function. Sequential analysis offers a possible alternative or complementary solution, where allocation is tuned mid-sample as yield estimates are developed.

We noted in Section 2.4 that the recall estimator is biased, and that the bias can be severe and positive, particularly where the prevalence of the unretrieved segment is low, and unretrieved sample size is small relative to prevalence. There is long-standing literature on unbiased ratio estimators [Hartley and Ross 1954; Al-Jararha 2008]. An unbiased estimator of recall is likewise desirable.

It was also noted in Section 2.4 that the sampling distribution of recall is in general non-symmetric, and especially so for extreme prevalences in the lower stratum. For the similarly non-symmetric sampling distribution of the binomial proportion, a logit-transformed two-tailed Wald interval on $\logit \pi$ of $\ln p/q \pm z/\sqrt{npq}$ has been found to give coverage characteristics similar to (though slightly wider than) the Wilson interval [Newcombe 2001]; the one-tailed variant also gives coverage similar to the Wilson [Liu and Kott 2009]. Applying such a transform to the normal approximation interval for recall might correct its spurious symmetry.⁹ The transformed interval, however, would still mishandle extreme cases; for instance, when estimated recall is 0 or 1, then the logit-transformed interval estimate is $[0, 0]$ and $[1, 1]$, respectively.

This article has not considered the thorny issue of assessor error. Such error has a particularly strong impact on low-prevalence corpora; even a low false positive can drastically inflate estimated yield. By sampling primary assessments for checking by a more authoritative source, the error rate can be estimated and corrected for [Webber et al. 2010]. The error-rate estimate, though, will have its own sampling error; and the more authoritative source, even if available, might still make mistakes. More complex approaches might overlap multiple assessors, to verify dubious assessments and estimate assessor reliability. The Bayesian beta-binomial model offers a flexible foundation on which to extend such models.

Only simple and stratified random sampling have been considered in this article. A more general approach is unequal sampling, in which each document is assigned its own inclusion probability, based on its probability of relevance and its weight in the evaluation metric. Unequal sampling has been applied in the Ad Hoc Task of the TREC Legal Track [Tomlinson et al. 2007; Hedin et al. 2009], and methods for unequal sampling and point (though not interval) estimation in ranked retrieval evaluation have been developed [Aslam et al. 2006; Aslam and Pavlu 2008]. Calculating variance for unequal sampling schemes requires calculating joint inclusion probabilities for sampled pairs. A simple sampling method for which joint inclusion probabilities are known is Sunter sampling [Sunter 1977; Särndal et al. 1992]; as it happens, the sampling scheme using in the Legal Ad Hoc Task is equivalent to Sunter sampling. Such variance estimators, however, will run into the same problem for low or zero-prevalence samples as simple random sampling, and in any case rely upon the normal or some similar approximation to turn into confidence intervals. Bayesian methods are also more complicated to use with unequal sampling, since we are no longer estimating a distribution over a single parameter (such as the proportion relevant in a segment), but require a more complex model linking probability of relevance to rank and other evidence [Carterette 2007].

Confidence intervals on precision in simple random sampling are straightforward cases of sampling for a binomial population, and can be dealt with using the Wilson interval described in Section 2.6, or the Jeffreys interval described in Section 2.7. Where stratified sampling is used, it needs to be determined whether the normal approximation used by Tomlinson et al. [2007] or the effective sample size approach described in Liu and Kott [2009] gives more reliable results. For both types of sampling, simple and stratified, other methods should be compared against the beta-binomial posterior approach described in this article, particularly where sample size makes up a substan-

⁹I owe this suggestion to Dave Lewis.

tial proportion of any stratum. Unlike precision, intervals for the F score are at least as complicated as for recall, as the F score is made up of recall and precision. The beta-binomial posterior method therefore seems preferable for F score intervals, though the degree of its preferability requires demonstration.

This article has considered two-sided confidence intervals. In some practical applications, particularly those of auditing or quality assurance, one-sided intervals are of more interest, in particular to set a lower bound on recall or precision. An interval method may obtain good two-sided coverage by having a narrow gap on one tail being compensated by a wide gap on the other; such a method unmodified does not produce a reliable one-sided interval. Liu and Kott [2009] examine one-sided intervals for a binomial population, under both simple and stratified random sampling, that correct bias in the normal approximation by taking higher terms of the Edgeworth expansion [Hall 1995]. These methods should be extended to the calculation of recall, and compared with the beta-binomial posterior; the good balance displayed by the latter in Figure 13 suggests that it may be competitive in this comparison.

Simple or stratified random sampling requires minimal assumptions, and as a result produces wide bounds on yield and recall. With stronger assumptions and models, tighter bounds can be achieved, at the cost of shifting the uncertainty to the validity of the assumptions. For instance, Kantor et al. [1999] propose a capture-recapture model, using the overlap between two runs to estimate yield, just as the individuals common to two separate trapping efforts estimate animal numbers in the wild. In its simplest form, this requires the unrealistic assumption that the two runs are independent, but a model that accounts for run correlation could be developed. Similarly, Zobel [1998] fits and extrapolates a model of proportion relevant at depth to ranked retrievals. Extending to depth one million a curve that has been observed to depth one hundred requires some courage, but the approach could be combined with sampling and other forms of evidence. As with sampling and confidence intervals, the task will be to provide not only a point estimate, but also some measure of that estimate's reliability.

Whatever estimation methods are used, and even if assessor error is excluded, bounds on recall for large collections with low prevalence and constrained assessment budgets will generally be wide, sometimes so wide as to appear unhelpful to the evaluator. But this is no reason to neglect the calculation and reporting of such bounds; quite the reverse. Where uncertainty is high, there is all the more reason to reveal that fact. Only once that is recognized can the effort be made to bring that uncertainty within tolerable limits.

ACKNOWLEDGMENTS

The author thanks Doug Oard, Dave Lewis, and Eric Slud for their perceptive suggestions; and Mossaab Bagdouri for observing the anomalous behaviour of the most-conservative prior for sample size of 1.

REFERENCES

- AGRESTI, A. AND CAFFO, B. 2000. Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures. *The American Statistician* 54, 4, 280–288.
- AGRESTI, A. AND COULL, B. A. 1998. Approximate is better than “exact” for interval estimation of binomial proportions. *The American Statistician* 52, 2, 119–126.
- AL-JARARHA, J. 2008. Unbiased ratio estimation for finite populations. Ph.D. thesis, Colorado State University.
- ASLAM, J. AND PAVLU, V. 2008. A practical sampling strategy for efficient retrieval evaluation. Tech. rep., Northeastern University.
- ASLAM, J., PAVLU, V., AND YILMAZ, E. 2006. A statistical method for system evaluation using incomplete judgments. In *Proc. 29th Annual International ACM SIGIR Conference on Research and Development*

- in *Information Retrieval*, S. Dumais, E. Efthimiadis, D. Hawking, and K. Järvelin, Eds. Seattle, Washington, USA, 541–548.
- BERGER, J. O., BERNARDO, J. M., AND SUN, D. 2008. Objective priors for discrete parameter spaces. Tech. rep., Duke University.
- BOLSTAD, W. M. 2007. *Introduction to Bayesian Statistics*. John Wiley & Sons.
- BROWN, L. D., CAI, T. T., AND DASGUPTA, A. 2001. Interval estimation for a binomial proportion. *Statistical Science* 18, 2, 101–133.
- BUCKLAND, S. T. 1984. Monte Carlo confidence intervals. *Biometrics* 40, 3, 811–817.
- CAI, T. T. 2005. One-sided confidence intervals in discrete distributions. *Journal of Statistical Planning and Inference* 131, 1, 63–88.
- CARTERETTE, B. 2007. Robust test collections for retrieval evaluation. In *Proc. 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, C. L. A. Clarke, N. Fuhr, N. Kando, W. Kraaij, and A. de Vries, Eds. Amsterdam, the Netherlands, 55–62.
- CHEN, M.-H. AND SHAO, Q.-M. 1999. Monte Carlo estimation of bayesian credible and HPD intervals. *Journal of Computational and Graphical Statistics* 8, 1, 69–92.
- CHENG, R. C. H. 1978. Generating beta variates with nonintegral shape parameters. *Communications of the ACM* 21, 4, 317–322.
- CLOPPER, C. J. AND PEARSON, E. S. 1934. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* 26, 4, 404–413.
- COCHRAN, W. G. 1977. *Sampling Techniques* 3rd Ed. John Wiley & Sons.
- DUTKA, J. 1984. The early history of the hypergeometric function. *Archive for History of Exact Sciences* 31, 1, 15–34.
- DYER, D. AND CHIOU, P. 1984. An information-theoretic approach to incorporating prior information in binomial sampling. *Communications in Statistics: Theory and Methods* 13, 17, 2051–2083.
- DYER, D. AND PIERCE, R. L. 1993. On the choice of the prior distribution in hypergeometric sampling. *Communications in Statistics: Theory and Methods* 22, 8, 2125–2146.
- EFRON, B. AND TIBSHIRANI, R. 1993. *An Introduction to the Bootstrap*. Chapman and Hall/CRC.
- FELLER, W. 1945. On the normal approximation to the binomial distribution. *The Annals of Mathematical Statistics* 16, 4, 319–329.
- GELMAN, A., CARLIN, J. B., STERN, H. S., AND RUBIN, D. B. 2004. *Bayesian Data Analysis* 2nd Ed. Chapman and Hall/CRC.
- GREENLAND, S. 2001. Agresti, A., and Caffo, B., “Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures,” *The American Statistician*, 54, 280–288: Comment by Greenland and reply. *The American Statistician* 55, 2, 172.
- GUENTHER, W. C. 1971. Unbiased confidence intervals. *The American Statistician* 25, 1, 51–53.
- HALL, P. 1982. Improving the normal approximation when constructing one-sided confidence intervals for binomial or Poisson parameters. *Biometrika* 69, 3, 647–652.
- HALL, P. 1995. *The Bootstrap and Edgeworth Expansion*. Springer.
- HARTLEY, H. O. AND ROSS, A. 1954. Unbiased ratio estimators. *Nature* 174, 270–271.
- HEDIN, B., TOMLINSON, S., BARON, J. R., AND OARD, D. W. 2009. Overview of the TREC 2009 legal track. In *Proc. 18th Text REtrieval Conference*, E. Voorhees and L. P. Buckland, Eds. Gaithersburg, Maryland, USA, 1:4:1–40. NIST Special Publication 500-278.
- JEFFREYS, H. 1946. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences* 186, 1007, 453–461.
- KANTOR, P., KIM, M.-H., IBRAEV, U., AND ATASOY, K. 1999. Estimating the number of relevant documents in enormous collections. In *Proc. ASIS Annual Meeting*. 507–514.
- KOOPMAN, P. A. R. 1984. Confidence intervals for the ratio of two binomial proportions. *Biometrics* 40, 2, 513–517.
- LEHMANN, E. L. AND CASELLA, G. 1998. *Theory of Point Estimation*. Springer.
- LEHMANN, E. L. AND ROMANO, J. P. 2005. *Testing Statistical Hypotheses* 3rd Ed. Springer.
- LIU, Y. K. AND KOTT, P. S. 2009. Evaluating alternative one-sided coverage intervals for a proportion. *Journal of Official Statistics* 25, 4, 569–588.
- NEWCOMBE, R. G. 2001. Logit confidence intervals and the inverse sinh transformation. *The American Statistician* 55, 3, 200–202.
- NEYMAN, J. 1935. On the problem of confidence intervals. *The Annals of Mathematical Statistics* 6, 3, 111–116.

- NICHOLSON, W. L. 1956. On the normal approximation to the hypergeometric distribution. *The Annals of Mathematical Statistics* 27, 2, 471–483.
- OARD, D. W., HEDIN, B., TOMLINSON, S., AND BARON, J. R. 2008. Overview of the TREC 2008 legal track. In *Proc. 17th Text REtrieval Conference*, E. Voorhees and L. P. Buckland, Eds. Gaithersburg, Maryland, USA, 3:1–45. NIST Special Publication 500-277.
- SÄRNDAL, C.-E., SWENSSON, B., AND WRETMAN, J. 1992. *Model assisted survey sampling*. Springer-Verlag.
- SIMEL, D. L., SAMSA, G. P., AND MATCHAR, D. B. 1991. Likelihood ratios with confidence: sample size estimation for diagnostic test studies. *Journal of Clinical Epidemiology* 44, 8, 763–770.
- SMITHSON, M. 2002. *Confidence intervals*. Sage Publications.
- SUNTER, A. B. 1977. List sequential sampling with equal or unequal probabilities without replacement. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 26, 3, 261–268.
- TAYLOR, J. R. 1997. *Introduction to error analysis* 2nd Ed. University Science Books.
- THOMPSON, S. K. 2002. *Sampling* 2nd Ed. John Wiley & Sons, New York.
- TOMLINSON, S., OARD, D. W., BARON, J. R., AND THOMPSON, P. 2007. Overview of the TREC 2007 legal track. In *Proc. 16th Text REtrieval Conference*, E. Voorhees and L. P. Buckland, Eds. Gaithersburg, Maryland, USA, 5:1–34. NIST Special Publication 500-274.
- WEBBER, W., OARD, D. W., SCHOLER, F., AND HEDIN, B. 2010. Assessor error in stratified evaluation. In *Proc. 19th ACM International Conference on Information and Knowledge Management*. Toronto, Canada, 539–548.
- ZOBEL, J. 1998. How reliable are the results of large-scale information retrieval experiments? In *Proc. 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, and J. Zobel, Eds. Melbourne, Australia, 307–314.

Received ; revised ; accepted